VARIATIONAL KULLBACK-LEIBLER DIVERGENCE FOR HIDDEN MARKOV MODELS

John R. Hershey, Peder A. Olsen, Steven J. Rennie

IBM Thomas J. Watson Research Center

ABSTRACT

Divergence measures are widely used tools in statistics and pattern recognition. The Kullback-Leibler (KL) divergence between two hidden Markov models (HMMs) would be particularly useful in the fields of speech and image recognition. Whereas the KL divergence is tractable for many distributions, including gaussians, it is not in general tractable for mixture models or HMMs. Recently, variational approximations have been introduced to efficiently compute the KL divergence and Bhattacharyya divergence between two mixture models, by reducing them to the divergences between the mixture components. Here we generalize these techniques to approach the divergence between HMMs using a recursive backward algorithm. Two such methods are introduced, one of which yields an upper bound on the KL divergence, the other of which yields a recursive closed-form solution. The KL and Bhattacharyya divergences, as well as a weighted edit-distance technique, are evaluated for the task of predicting the confusability of pairs of words.

Index Terms: Kullback-Leibler divergence, variational methods, mixture models, hidden Markov models (HMMs), weighted edit distance, Bhattacharyya divergence.

1. INTRODUCTION

The Kullback-Leibler (KL) divergence, also known as the *relative entropy*, between two probability density functions $f(x)$ and $g(x)$,

$$D(f\|g) \stackrel{\text{def}}{=} \int f(x) \log \frac{f(x)}{g(x)} \, \mathrm{dx}, \tag{1}$$

is commonly used in statistics as a measure of similarity between two density distributions [1]. The KL divergence satisfies three *divergence properties*:

1. Self similarity: $D(f\|f) = 0$.
2. Self identification: $D(f\|g) = 0$ only if $f = g$.
3. Positivity: $D(f\|g) \geq 0$ for all f, g.

The KL divergence is used in many aspects of speech and image recognition, such as determining if two acoustic models are similar, [2], measuring how confusable two words or hidden Markov models (HMMs) are, [3, 4, 5], computing the best match using pixel distribution models [6], clustering of models, and optimization by minimizing or maximizing the divergence between distributions.

The KL divergence has a closed form expression for many probability densities. For two gaussians, f and g, it reduces to the well-known expression,

$$D(f\|g) = \frac{1}{2}\Big[\log \frac{|\Sigma_g|}{|\Sigma_f|} + \mathrm{Tr}[\Sigma_g^{-1}\Sigma_f] - d + (\mu_f - \mu_g)^T \Sigma_g^{-1} (\mu_f - \mu_g)\Big]. \tag{2}$$

{jrhershe,pederao,sjrennie }@us.ibm.com

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. These densities are defined as $f(x) \stackrel{\text{def}}{=} \exp(\theta_f^T \phi(x))/z(\theta_f)$, where θ_f is a vector of parameters, and $z(\theta_f) = \int \exp(\theta_f^T \phi(x)) \, \mathrm{dx}$, and $\phi(x)$ is a vector-valued function of x [7]. This formulation makes the KL divergence between two such densities surprisingly simple:

$$D(f\|g) \quad = \quad \log \frac{z(\theta_f)}{z(\theta_g)} + (\theta_f - \theta_g)^T E_f \phi(x), \tag{3}$$

which requires only that $E_f \phi(x)$ be known [8].

In general, however, for more complex distributions such as *mixture models* and hidden Markov models, the integral involves the logarithm of sums of component densities, and no such simple expression exists. In the following sections we review two variational approximations to the KL divergence between two mixture models. Throughout the paper we use the example of *gaussian mixture models* (GMMs), and HMMs with gaussian mixtures as observation models, although the same techniques directly apply to any densities for which we can compute the KL divergences between pairs of mixture components.

For an observation of a given sequence length an HMM can be construed as a mixture model in which each HMM state sequence is a mixture component. In theory, the variational approximations for the KL divergence between two mixture models directly carries over to HMMs in this sense. However, the direct application of the variational approximation yields one set of state sequences inside the logarithm, and another outside the logarithm. This prevents us from using a recursive formulation to sum over the exponential number of pairs of state sequences generated by typical HMMs.

Therefore we derive a new variational approximation that is amenable to standard forward and backward algorithms. The variational approximations contain variational parameters that serve to associate sequences of one HMM with similar sequences of the other. We constrain these parameters by factorizing them into a Markov chain, which allows us to recursively solve for the variational parameters and evaluate the approximation.

One weakness of the KL divergence between HMMs is that, in many common cases, the divergence becomes infinite. If the HMM f generates sequences of lengths that the HMM g cannot generate, then the KL divergence is infinite. Such is the case, for instance, in left-to-right models where g is longer than f, despite the fact that the precise length of such models may be an artifact of the phonetic system of the recognizer, rather than an important modeling assumption. In [9], this was addressed by connecting the final state to the initial state to make the HMM ergodic, and substituting the KL divergence rate for the KL divergence. Here we propose methods for approximating KL divergences of non-ergodic HMMs, and instead we consider symmetric versions of the KL divergence that yields finite meaningful values. We consider two symmetrized versions of

the KL divergence:

$$\begin{aligned} D_{\min}(f,g) &= \min\{D(f\|g), D(g\|f)\} \\ D_{\text{resistor}}(f,g) &= \left(D^{-1}(f\|g) + D^{-1}(g\|f)\right)^{-1} \end{aligned}$$

The *resistor average* symmetrized KL divergence was first introduced in [10]. This problem can also be addressed by computing the KL divergence over the intersection of the sets of sequence lengths allowed by the HMMs.

In addition to the KL divergence, there exist other useful measures of dissimilarity between distributions. In particular, the Bhattacharyya divergence

$$D_B(f,g) \stackrel{\text{def}}{=} -\log \int \sqrt{f(x)g(x)}\, \mathrm{dx}, \tag{4}$$

is closely related to the KL divergence and can be used to bound the *Bayes error*, $B_e(f,g) \stackrel{\text{def}}{=} \frac{1}{2}\int \min(f(x), g(x))\ \mathrm{dx} \leq \frac{1}{2}e^{-D_B(f,g)}$. The Bhattacharyya divergence is symmetric, and has the advantage that it does not diverge to infinity for HMMs which support different sets of sequence lengths, so long as both HMMs can generate some sequences of the same length. The variational approximations for the KL divergence can also be applied to compute a variational approximation to the Bhattacharyya divergence, without factorizing the variational parameters. It also turns out that the Bhattacharyya divergence is closely related to a heuristic method known as the weighted edit distance, which we include here in our experiments. To validate these approaches we compare numerical predictions with empirical word confusability measurements (i.e., word substitution error rates) from a speech recognizer.

2. VARIATIONAL METHODS FOR MIXTURE MODELS

In [11] variational methods were introduced that allow the KL divergence to be approximated for mixture models. Without loss of generality, we consider the case where f and g are gaussian mixture models, with marginal densities of $x \in \mathbb{R}^d$ under f and g given by

$$\begin{aligned} f(x) &= \textstyle\sum_a \pi_a \mathcal{N}(x; \mu_a; \Sigma_a) \\ g(x) &= \textstyle\sum_b \omega_b \mathcal{N}(x; \mu_b; \Sigma_b). \end{aligned} \tag{5}$$

Here 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 variance Σ_a. We 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 $D(f\|g)$ will make use of the KL-divergence between individual components, which we thus write as $D(f_a\|g_b)$.

The Variational Approximation for Mixture Models: A variational lower bound to the likelihood was introduced in [11]. We define variational parameters $\phi_{b|a} > 0$ such that $\sum_b \phi_{b|a} = 1$. By Jensen's inequality we have

$$\begin{aligned} L(f\|g) &\stackrel{\text{def}}{=} \int f(x) \log g(x)\ \mathrm{dx} \\ &= \sum_a \pi_a \int f_a(x) \log \sum_b \phi_{b|a} \frac{\omega_b g_b(x)}{\phi_{b|a}}\ \mathrm{dx} \\ &\geq \sum_a \pi_a \sum_b \phi_{b|a} \left(\log \frac{\omega_b}{\phi_{b|a}} + L(f_a\|g_b)\right) \quad (6) \\ &\stackrel{\text{def}}{=} \mathcal{L}_\phi(f\|g), \quad (7) \end{aligned}$$

where $L(f_a\|g_b) \stackrel{\text{def}}{=} \int f_a(x) \log g_b(x)\ \mathrm{dx}$. Since this is a lower bound on $L(f\|g)$, we get the best bound by maximizing $\mathcal{L}_\phi(f\|g)$ with respect to ϕ. If we define $D_{\text{VA}}(f\|g) = \mathcal{L}_{\hat{\psi}}(f\|f) - \mathcal{L}_{\hat{\phi}}(f\|g)$ and substitute the optimal variational parameters, $\hat{\phi}_{b|a}$ and $\hat{\psi}_{a'|a}$, the result simplifies to

$$D_{\text{VA}}(f\|g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a\|f_{a'})}}{\sum_b \omega_b e^{-D(f_a\|g_b)}}. \tag{8}$$

$D_{\text{VA}}(f\|g)$ satisfies the similarity property, but it does not in general satisfy the positivity property. Note that this variational approximation is the difference of two bounds, and hence is not itself a bound. In terms of accuracy, however, it performs somewhat better than the bound described below perhaps because some of the error cancels out in the subtraction, as shown in [11].

The Variational Bound for Mixture Models: A direct upper bound on the divergence is also introduced in [11] for mixture models. We define the variational parameters $\phi_{ab} \geq 0$ and $\psi_{ab} \geq 0$ satisfying the constraints $\sum_b \phi_{ab} = \pi_a$ and $\sum_a \psi_{ab} = \omega_b$. Using the variational parameters we may write

$$\begin{aligned} f &= \textstyle\sum_a \pi_a f_a = \sum_{ab} \phi_{ba} f_a \\ g &= \textstyle\sum_b \omega_b g_b = \sum_{ab} \psi_{ab} g_b. \end{aligned} \tag{9}$$

With this notation we use Jensen's inequality to obtain an upper bound of the KL divergence as follows

$$\begin{aligned} D(f\|g) &= \int f \log(f/g) \\ &= -\int f \log \left(\sum_{ab} \frac{\psi_{ab} g_b}{\phi_{ab} f_a}\ \frac{\phi_{ab} f_a}{f}\right) \mathrm{dx} \\ &\leq \sum_{ab} \phi_{ab} \int f_a \log \left(\frac{\phi_{ab} f_a}{\psi_{ab} g_b}\right) \mathrm{d}x \quad (10) \\ &\stackrel{\text{def}}{=} \mathcal{D}_{\phi\psi}(f\|g). \end{aligned}$$

The best possible upper bound can be attained by finding the variational parameters $\hat{\phi}$ and $\hat{\psi}$ that minimize $\mathcal{D}_{\phi\psi}(f\|g)$. The problem is convex in ϕ as well as in ψ so we can fix one and optimize for the other. Fixing ϕ the optimal value for ψ is seen to be

$$\psi_{ab} = \frac{\omega_b \phi_{ab}}{\sum_{a'} \phi_{a'b}}. \tag{11}$$

Similarly, fixing ψ the optimal value for ϕ is

$$\phi_{ab} = \frac{\pi_a \psi_{ab} e^{-D(f_a\|g_b)}}{\sum_{b'} \psi_{ab'} e^{-D(f_a\|g_{b'})}}. \tag{12}$$

At each iteration step the upper bound $\mathcal{D}_{\phi\psi}(f\|g)$ is lowered, and we refer to the convergent as $D_{\text{VB}}(f\|g)$. Since any zeros in ϕ and ψ are fixed under the iteration we recommend starting with $\phi_{ab} = \psi_{ab} = \pi_a \omega_b$. In practice it converges sufficiently in a few iterations [11]. This iterative scaling scheme is of the same type as the Blahut-Arimoto algorithm for computing the channel capacity and also arises in maximum entropy models (see [11] for references).

3. HIDDEN MARKOV MODELS

To formulate the KL 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. To this end the HMM must terminate the sequence when it transitions to a special final state. For an HMM, f, emitting an observation sequence of length n, as

$a_{1:n} \stackrel{\text{def}}{=} (a_1, \ldots, a_n)$ 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 $x_{1:n} \stackrel{\text{def}}{=} (x_1, \ldots, x_n)$ be a sequence of observations, with $x_t \in \mathbb{R}^d$. For the observations we use the shorthand $f_{a_t}(x_t) \stackrel{\text{def}}{=} \mathcal{N}(x_t; \mu_{a_t}, \Sigma_{a_t})$. We also define non-emitting initial and final state values (i.e., not random variables) $_{\mathcal{I}}$, and $_{\mathcal{F}}$. The state sequence probabilities are thus formulated as a Markov chain $\pi_{a_{1:n}} \stackrel{\text{def}}{=} \pi_{a_1|\mathcal{I}}\, \pi_{\mathcal{F}|a_n} \prod_{t=2}^{n} \pi_{a_t|a_{t-1}}$, 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 $t \geq 2$. It bears emphasizing here that the transitions to emitting states do not in general sum to one (i.e., $\sum_{a_t} \pi_{a_t|a_{t-1}} \leq 1$), because there may also be a transition to the non-emitting final state. Hence it is as if the HMM is gradually leaking probability away to paths which terminate before the path in question. This allows the HMM to describe a distribution over all sequence lengths. In general, the transition to the final state can occur at any time; however, for a given sequence length n, we only consider paths that reach the final state after exactly n observations. The density assigned to signals of particular length can thus be written:

$$\begin{aligned} f(x_{1:n}) &= \sum_{a_{1:n}} \pi_{a_{1:n}} f_{a_{1:n}}(x_{1:n}) \\ &= \sum_{a_{1:n}} \pi_{a_1|\mathcal{I}} \pi_{\mathcal{F}|a_n} f_{a_1}(x_1) \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} f_{a_t}(x_t). \end{aligned}$$

The probability of a particular sequence length n is $p_f(n) = \int f(x_{1:n})\, \mathrm{dx}_{1:n} = \sum_{a_{1:n}} \pi_{a_1|\mathcal{I}} \pi_{\mathcal{F}|a_n} \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} \leq 1$. Since the set of all sequences is $\mathbf{x} \in \cup_{n=1}^{\infty} \mathbb{R}^{n \times d}$, the integration over all sequences is perhaps an unfamiliar operation. It amounts to separately integrating over sequences of each length and then summing over the individual results. To see that f is a distribution over all sequences it is enough to verify that indeed

$$\int f(\mathbf{x}) \mathrm{d}\mathbf{x} = \sum_{n=1}^{\infty} \int f(x_{1:n})\, \mathrm{dx}_{1:n} = \sum_{n=1}^{\infty} p_f(n) = 1. \quad (13)$$

4. THE VARIATIONAL APPROXIMATION FOR HMMS

We extend the variational approximation for mixture models to HMMs by defining variational parameters in the form of a conditional Markov chain, $\phi_{b_{1:n}|a_{1:n}} \stackrel{\text{def}}{=} \phi_{b_1|a_1} \prod_{t=2}^{n} \phi_{b_t|a_t b_{t-1}}$ where $\sum_{b_1} \phi_{b_1|a_1} = 1$ and $\sum_{b_t} \phi_{b_t|a_t b_{t-1}} = 1$, so that $\sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} = 1$.
For a given sequence length n, by Jensen's inequality we have

$$\begin{aligned} L_n(f\|g) &\stackrel{\text{def}}{=} \int f(x_{1:n}) \log g(x_{1:n})\, \mathrm{dx}_{1:n} \\ &\geq \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} \log \frac{\omega_{b_{1:n}} e^{L(f_{a_{1:n}}\|g_{b_{1:n}})}}{\phi_{b_{1:n}|a_{1:n}}} \\ &\stackrel{\text{def}}{=} \mathcal{L}_\phi(f\|g), \end{aligned} \quad (14)$$

where $L(f_{a_{1:n}}\|g_{b_{1:n}}) \stackrel{\text{def}}{=} \int f_{a_{1:n}}(x_{1:n}) \log g_{b_{1:n}}(x_{1:n})\ \ \mathrm{dx}_{1:n}$. Note that

$$L(f_{a_{1:n}}\|g_{b_{1:n}}) = \sum_{t=1}^{n} L(f_{a_t}\|g_{b_t}) \quad (15)$$

where $L(f_{a_t}\|g_{b_t}) \stackrel{\text{def}}{=} \int f_{a_t}(x_t) \log g_{b_t}(x_t)\, \mathrm{dx}_t$
Since this is a lower bound on $\mathcal{L}(f\|g)$, we get the best bound by maximizing $\mathcal{L}_\phi(f\|g)$ with respect to $\phi_{b_{1:n}|a_{1:n}}$. To do so, we first expand the objective function into a recursive formula, by pulling earlier terms out of the sums over later variables.

$$\begin{aligned} \mathcal{L}_\phi(f\|g) = &\sum_{a_1} \pi_{a_1|\mathcal{I}} \sum_{b_1|\mathcal{I}} \phi_{b_1|a_1} \Big(\\ &\qquad (\sum_{a_{2:n}} \pi_{a_{2:n}|a_1}) \log \frac{\omega_{b_1} e^{L(f_{a_1}\|g_{b_1})}}{\phi_{b_1|a_1}} \\ &+ \sum_{a_2} \pi_{a_2|a_1} \sum_{b_2} \phi_{b_2|a_2 b_1} \Big(\\ &\qquad (\sum_{a_{3:n}} \pi_{a_{3:n}|a_2}) \log \frac{\omega_{b_2|b_1} e^{L(f_{a_2}\|g_{b_2})}}{\phi_{b_2|a_2 b_1}} \\ &+ \ldots \\ &+ \sum_{a_{n-1}} \pi_{a_{n-1}|a_{n-2}} \sum_{b_{n-1}} \phi_{b_{n-1}|a_{n-1} b_{n-2}} \Big(\\ &\qquad (\sum_{a_n} \pi_{a_n|a_{n-1}} \pi_{\mathcal{F}|a_n}) \log \frac{\omega_{b_{n-1}|b_{n-2}} e^{L(f_{a_{n-1}}\|g_{b_{n-1}})}}{\phi_{b_{n-1}|a_{n-1} b_{n-2}}} \\ &+ \sum_{a_n} \pi_{a_n|a_{n-1}} \sum_{b_n} \phi_{b_n|a_n b_{n-1}} \Big(\\ &\qquad \pi_{\mathcal{F}|a_n} \log \frac{\omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} e^{L(f_{a_n}\|g_{b_n})}}{\phi_{b_n|a_n b_{n-1}}} \Big) \ldots \Big) \Big), \end{aligned} \quad (16)$$

where we can use the following recursion to compute the nested sums over the priors

$$\begin{aligned} p_{n-t}(a_t) &\stackrel{\text{def}}{=} \sum_{a_{t+1:n}} \pi_{a_{t+1:n}|a_t} = \sum_{a_{t+1:n}} \pi_{\mathcal{F}|a_n} \prod_{\tau=t+1}^{n} \pi_{a_\tau|\tau-1} \\ &= \sum_{a_{t+1}} \pi_{a_{t+1}|a_t} p_{n-t-1}(a_{t+1}) \end{aligned} \quad (17)$$

is the probability that a sequence in state a_t will terminate in $n-t$ steps. The recursion terminates with $p_0(a_n) \stackrel{\text{def}}{=} \pi_{\mathcal{F}|a_n}$.
Then we can write (16) recursively as

$$\begin{aligned} \mathcal{L}_t^\phi(a_{t-1}, b_{t-1}) \stackrel{\text{def}}{=} &\sum_{a_t} \pi_{a_t|a_{t-1}} \sum_{b_t} \phi_{b_t|a_t b_{t-1}} \Big(\\ &\qquad p_{n-t}(a_t) \log \frac{\omega_{b_t|b_{t-1}} e^{L(f_{a_t}\|g_{b_t})}}{\phi_{b_t|a_t b_{t-1}}} + \mathcal{L}_{t+1}^\phi(a_t, b_t) \Big), \end{aligned}$$

beginning the recursion with

$$\begin{aligned} \mathcal{L}_n^\phi(a_{n-1}, b_{n-1}) = &\sum_{a_n} \pi_{a_n|a_{n-1}} \sum_{b_n} \phi_{b_n|a_n b_{n-1}} \Big(\\ &\qquad p_0(a_n) \log \frac{\omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} e^{L(f_{a_n}\|g_{b_n})}}{\phi_{b_n|a_n b_{n-1}}} \Big), \end{aligned}$$

and terminating it with

$$\begin{aligned} \mathcal{L}_\phi(f\|g) = &\sum_{a_1} \pi_{a_1|\mathcal{I}} \sum_{b_1} \phi_{b_1|a_1} \Big(\\ &\qquad p_{n-1}(a_1) \log \frac{\omega_{b_1|\mathcal{I}} e^{L(f_{a_1}\|g_{b_1})}}{\phi_{b_1|a_1}} + \mathcal{L}_2^\phi(a_1, b_1) \Big). \end{aligned}$$

Note that $\mathcal{L}_t^{\phi}(a_{t-1}, b_{t-1})$ is the only term containing $\phi_{b_t|a_t b_{t-1}}$, so the derivative is

$$\frac{\partial \mathcal{L}_{\phi}(f\|g)}{\partial \phi_{b_t|a_t b_{t-1}}} = \tilde{\phi}_{b_{t-1}} \tilde{\pi}_{a_t} \Big(p_{n-t}(a_t) \log \frac{\omega_{b_t|b_{t-1}} e^{L(f_{a_t}\|g_{b_t})}}{\phi_{b_t|a_t b_{t-1}}} + \mathcal{L}_{t+1}^{\phi}(a_t, b_t) - p_{n-t}(a_t) \Big).$$

Where $\tilde{\phi}_{b_{t-1}} \tilde{\pi}_{a_t}$ are some priors that are independent of b_t. Equating to zero and solving for $\phi_{b_t|a_t b_{t-1}}$ yields

$$\hat{\phi}_{b_t|a_t b_{t-1}} = \frac{\omega_{b_t|b_{t-1}} e^{L(f_{a_t}\|g_{b_t})} e^{\mathcal{L}_{t+1}^{\phi}(a_t,b_t)/p_{n-t}(a_t)}}{\sum_{b_t} \omega_{b_t|b_{t-1}} e^{L(f_{a_t}\|g_{b_t})} e^{\mathcal{L}_{t+1}^{\phi}(a_t,b_t)/p_{n-t}(a_t)}}.$$

The variational parameters for the end points are:

$$\hat{\phi}_{b_n|a_n b_{n-1}} = \frac{\omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} e^{L(f_{a_n}\|g_{b_n})}}{\sum_{b_n} \omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} e^{L(f_{a_n}\|g_{b_n})}}.$$

and

$$\hat{\phi}_{b_1|a_1} = \frac{\omega_{b_1|\mathcal{I}} e^{L(f_{a_1}\|g_{b_1})} e^{\mathcal{L}_2^{\phi}(a_1,b_1)/p_{n-1}(a_1)}}{\sum_{b_1} \omega_{b_1|\mathcal{I}} e^{L(f_{a_1}\|g_{b_1})} e^{\mathcal{L}_2^{\phi}(a_1,b_1)/p_{n-1}(a_1)}}.$$

Substituting back in, we can simplify and eliminate the variational parameters altogether.

$$\mathcal{L}_t^{\hat{\phi}}(a_{t-1}, b_{t-1}) = \sum_{a_t} \pi_{a_t|a_{t-1}} \log \sum_{b_t} \omega_{b_t|b_{t-1}} e^{L(f_{a_t}\|g_{b_t})} e^{\mathcal{L}_{t+1}^{\hat{\phi}}(a_t,b_t)/p_{n-t}(a_t)}.$$

The recursion begins with

$$\mathcal{L}_n^{\hat{\phi}}(a_{n-1}, b_{n-1}) = \sum_{a_n} \pi_{a_n|a_{n-1}} \pi_{\mathcal{F}|a_n} \log \sum_{b_n} \omega_{b_n|b_{n-1}} \omega_{\mathcal{F}|b_n} e^{L(f_{a_n}\|g_{b_n})},$$

and ends with

$$L_{\mathrm{VA}}(f(x_{1:n})\|g(x_{1:n})) \stackrel{\text{def}}{=} \mathcal{L}_{\hat{\phi}}(f\|g) = \sum_{a_1} \pi_{a_1|\mathcal{I}} \log \sum_{b_1} \omega_{b_1|\mathcal{I}} e^{L(f_{a_1}\|g_{b_1})} e^{\mathcal{L}_2^{\hat{\phi}}(a_1,b_1)/p_{n-1}(a_1)}.$$

Thus we have a single-pass backward algorithm. The KL divergence is then approximated using

$$D_{\mathrm{VA}}(f\|g) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} L_{\mathrm{VA}}(f(x_{1:n})\|f(x_{1:n})) - L_{\mathrm{VA}}(f(x_{1:n})\|g(x_{1:n})),$$

which in practice is truncated to a finite series. Note that this sum can also be computed recursively by saving intermediate results.

In some situations a forward algorithm may be useful. In such cases an easy option is to make change of parameters to reverse the HMM itself, using time-dependent state transitions that condition on the future rather than the past. Reversing the HMM entails a forward algorithm to produce the reversed parameters. Then the above backward algorithms can be applied on the time-reversed HMM, yielding a second forward algorithm. The two forward algorithms can be done simultaneously, deriving the next parameters from the reversal algorithm just in time for the next forward step of the variational recursion.

5. THE VARIATIONAL BOUND FOR HMMS

To upper-bound the divergence we employ two variational parameters, again factorized into a Markov chain. A different upper bound is proposed in [12], in which the closest pair of paths is used instead of summing over all paths. In this section we define $c_t = (a_t, b_t)$ to simplify the notation. For HMMs we formulate the variational parameters as $\phi_{c_{1:n}} \stackrel{\text{def}}{=} \phi_{c_1|\mathcal{I}} \phi_{\mathcal{F}|c_n} \prod_{t=2:n} \phi_{c_t|c_{t-1}}$, and $\psi_{c_{1:n}} \stackrel{\text{def}}{=} \psi_{c_1|\mathcal{I}} \psi_{\mathcal{F}|c_n} \prod_{t=2:n} \phi_{c_t|c_{t-1}}$. We also have the constraints that $\sum_{b_t} \phi_{a_t b_t|a_{t-1} b_{t-1}} = \pi_{a_t|a_{t-1}}$ and $\sum_{a_t} \psi_{a_t b_t|a_{t-1} b_{t-1}} = \omega_{b_t|b_{t-1}}$. The variational parameters for the final state transitions are constrained to be $\phi_{\mathcal{F}|c_n} = \pi_{\mathcal{F}|a_n}$ $\psi_{\mathcal{F}|c_n} = \omega_{\mathcal{F}|b_n}$.

For the variational bound we have

$$D\left(f(x_{1:n})\|g(x_{1:n})\right) \le \mathcal{D}_{\phi\psi}(f\|g) \stackrel{\text{def}}{=} \sum_{c_{1:n}} \phi_{c_{1:n}} \left(\log \frac{\phi_{c_{1:n}}}{\psi_{c_{1:n}}} + D(f_{a_{1:n}}\|g_{b_{1:n}}) \right),$$

where

$$D(f_{a_{1:n}}\|g_{b_{1:n}}) \stackrel{\text{def}}{=} \int f_{a_{1:n}}(x_{1:n}) \log \frac{f_{a_{1:n}}(x_{1:n})}{g_{b_{1:n}}(x_{1:n})} \, \mathrm{dx}_{1:n}.$$

Note that, due to the conditional independence of the x_t given the a_t, we have

$$D(f_{a_{1:n}}\|g_{b_{1:n}}) = \sum_{t=1}^{n} D(f_{a_t}\|g_{b_t}), \tag{18}$$

where $D(f_{a_t}\|g_{b_t}) \stackrel{\text{def}}{=} \int f_{a_t}(x_t) \log \frac{f_{a_t}(x_t)}{g_{b_t}(x_t)} \, \mathrm{dx}_t$.

We unroll this in time as

$$\mathcal{D}_{\phi\psi}(f_{1:n}\|g_{1:n}) = \tag{19}$$
$$\sum_{c_1} \phi_{c_1|\mathcal{I}} \Big(\underset{1}{(} \sum_{c_{2:n}} \phi_{c_{2:n}|c_1}) \log \frac{\phi_{c_1|\mathcal{I}} e^{D(f_{a_1}\|g_{b_1})}}{\psi_{c_1|\mathcal{I}}}$$
$$+ \sum_{c_2} \phi_{c_2|c_1} \Big(\underset{2}{(} \sum_{c_{3:n}} \phi_{c_{3:n}|c_2}) \log \frac{\phi_{c_2|c_1} e^{D(f_{a_2}\|g_{b_2})}}{\psi_{c_2|c_1}} + \ldots +$$
$$\sum_{c_n} \phi_{c_n|c_{n-1}} \Big(\underset{n}{\phi_{\mathcal{F}|c_n}} \log \frac{\phi_{c_n|c_{n-1}} \phi_{\mathcal{F}|c_n}}{\psi_{c_n|c_{n-1}} \psi_{\mathcal{F}|c_n}} + D(f_{a_n}\|g_{b_n}) \Big) \cdots \Big) \Big).$$
$$\underset{2}{} \underset{1}{}$$

Because of the variational constraints we have the following equality

$$\sum_{c_{t+1:n}} \phi_{c_{t+1:n}|c_t} = \sum_{a_{t+1:n}} \sum_{b_{t+1:n}} \phi_{\mathcal{F}|c_n} \prod_{\tau=t}^{n-1} \phi_{c_{\tau+1}|c_\tau} \tag{20}$$
$$= \sum_{a_{t+1:n}} \pi_{\mathcal{F}|a_n} \prod_{\tau=t}^{n-1} \pi_{a_{\tau+1}|a_\tau} = \sum_{a_{t+1:n}} \pi_{a_{t+1:n}|a_t} = p_{n-t}(a_t).$$

Using (20) we can directly write the recursive form of (19) as

$$\mathcal{D}_t^{\phi\psi}(c_{t-1}) = \sum_{c_t} \phi_{c_t|c_{t-1}} \Big(p_{n-t}(a_t) \left[\log \frac{\phi_{c_t|c_{t-1}} e^{D(f_{a_t}\|g_{b_t})}}{\psi_{c_t|c_{t-1}}} \right] + \mathcal{D}_{t+1}^{\phi\psi}(c_t) \Big)$$

Beginning the recursion with

$$\mathcal{D}_n^{\phi\psi}(c_{n-1}) = \sum_{c_n} \phi_{\mathcal{F}|c_n}\phi_{c_n|c_{n-1}}\Bigg(\left[\log \frac{\phi_{\mathcal{F}|c_n}\phi_{c_n|c_{n-1}}e^{D(f_{a_n}\|g_{b_n})}}{\psi_{\mathcal{F}|c_n}\psi_{c_n|c_{n-1}}}\right]\Bigg)$$

and terminating it with

$$\mathcal{D}^{\phi\psi}(f\|g) = \sum_{c_1} \phi_{c_1|\mathcal{I}}\Bigg(p_{n-1}(a_1)\left[\log \frac{\phi_{c_1|\mathcal{I}}e^{D(f_{a_1}\|g_{b_1})}}{\psi_{c_1|\mathcal{I}}}\right] + \mathcal{D}_2^{\phi\psi}(c_1)\Bigg)$$

To optimize we must iterate between solving for $\phi_{c_t|c_{t-1}}$ and $\psi_{c_t|c_{t-1}}$, holding the other constant. The optimal value of $\phi_{c_t|c_{t-1}}$ given $\psi_{c_t|c_{t-1}}$ is

$$\hat{\phi}_{c_t|c_{t-1}} = \frac{\pi_{a_t|a_{t-1}}\psi_{c_t|c_{t-1}}e^{-D(f_{a_t}\|g_{b_t})-\mathcal{D}_{t+1}^{\phi\psi}(c_t)/p_{n-t}(a_t)}}{\sum_{b'_t}\psi_{a_tb'_t|a_tb_{t-1}}e^{-D(f_{a_t}\|g_{b'_t})-\mathcal{D}_{t+1}^{\phi\psi}(a_t,b'_t)/p_{n-t}(a_t)}}.$$

Similarly the optimal value for $\psi_{c_t|c_{t-1}}$ given $\phi_{c_t|c_{t-1}}$ is

$$\hat{\psi}_{c_t|c_{t-1}} = \frac{\omega_{b_t|b_{t-1}}p_{n-t}(a_t)\phi_{c_t|c_{t-1}}}{\sum_{a'_t}p_{n-t}(a'_t)\phi_{a'_tb_t|c_{t-1}}}.$$

The iteration can be done to convergence for each step in a backward algorithm. Analogously to the variational approximation we need corresponding starting and terminating iterations too.
Let $D_{VB}(f_{1:n}\|g_{1:n})$ be the convergent value of $\mathcal{D}_{\hat{\phi},\hat{\psi}}^{\phi\psi}$. Then $D_{VB}(f\|g) = \sum_{n=1}^{\infty} D_{VB}(f_{1:n}\|g_{1:n})$. Factoring the ϕ and ψ differently leads to a family of related approximations. Factorizations that constrain the variational parameters more will in general require less storage for variational parameters, and yield less accurate results. In addition, the values of the variational parameters can be constrained. Constraining them to be sparse leads to Viterbi-style dynamic programming algorithms for the KL divergence, which may have some computational advantages.

6. WEIGHTED EDIT DISTANCES

Various types of *weighted edit distances* have been applied to the task of estimating spoken word confusability, as discussed in [3] and [4]. A word is modeled in terms of a left–to–right HMM, see Fig. 1.

Fig. 1. 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. 2. This structure is similar to that used for acoustic perplexity [3] and the average divergence distance [4].
Weights are placed on the vertices that assign smaller values when the corresponding phoneme state models are more confusable. The

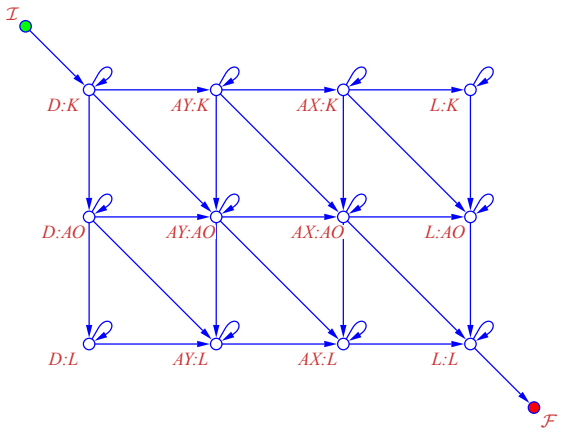

Fig. 2. Product HMM for the words *call* (`K AO L`) and *dial* (`D AY AX L`)

weighted edit distance (WED) is the shortest path (i.e., the Viterbi path) from the initial to the final node in the product graph.

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

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 $w_{f_{a_t},g_{b_t}}$ are dissimilarity measures between the acoustic models for each pair of HMM states. For the KL divergence WED, we define $w_{f_{a_t},g_{b_t}} \stackrel{\text{def}}{=} D(f_{a_t}\|g_{b_t})$, and for the Bhattacharyya WED, we define $w_{f_{a_t},g_{b_t}} \stackrel{\text{def}}{=} D_B(f_{a_t}\|g_{b_t})$. An interesting variation, which we call the *total weighted edit distance* TWED, is to sum over all paths and sequence lengths:

$$D_{\text{TWED}}(f,g) = -\log \sum_n \sum_{a_{1:n},b_{1:n}} e^{-C(a_{1:n},b_{1:n})}. \qquad (21)$$

That is, we sum over the similarities (probabilities), rather than the costs (negative log probabilities), since this corresponds to the interpretation as a product HMM.
It turns out that when we apply the variational techniques introduced above to the Bhattacharyya divergence, the resulting measure $D_B(f\|g)$ can be seen as a special case of the total weighted edit distance. These are formulated for mixture models in [13] and the same methods apply directly to HMMs. In addition, the TWED with Bhattacharyya weights, $w_{f_{a_t},g_{b_t}} \stackrel{\text{def}}{=} D_B(f_{a_t}\|g_{b_t})$ is in fact a simple Jensen's bound on the HMM Bhattacharyya divergence. Because the Bhattacharyya approximations and weighted edit distances are not in general zero for $f = g$, we subsequently normalize them using: $D_{\text{norm}}(f,g) = D(f,g) - \frac{1}{2}D(f,f) - \frac{1}{2}D(g,g)$, which improves the performance. The derivations of the variational Bhattacharyya divergence bounds and the details of their relationship to the weighted edit distances are beyond the scope of this paper and are to be published elsewhere.

7. WORD CONFUSABILITY EXPERIMENTS

In this section we briefly 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 $\mathrm{E}(w_1, w_2) \stackrel{\text{def}}{=} \frac{1}{2}\mathrm{P}(w_1|w_2) + \frac{1}{2}\mathrm{P}(w_2|w_1)$, 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 counts were too low. Continuous speech was used, so it is possible that some errors were due to mis-alignment.

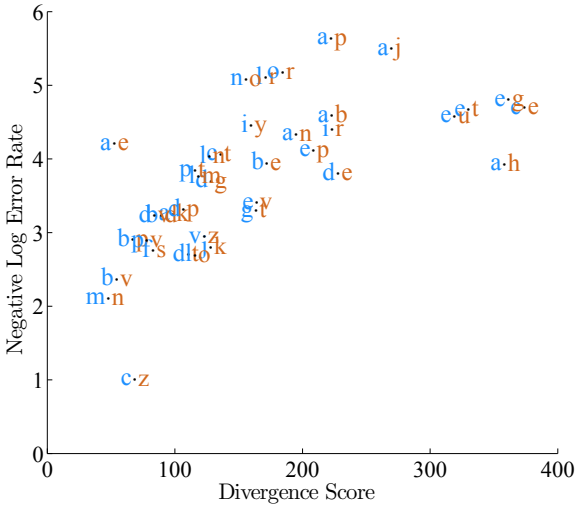

Fig. 3. The negative log error rate for all spelling word pairs compared to the variational HMM KL divergence.

Method	Score
VA Min KL Divergence	0.365
VA Resistor KL Divergence	0.433
MC 100K Min KL Divergence	0.450
MC 100K Resistor KL Divergence	0.442
KL Divergence Weighted Edit Distance	0.571
Bhattacharyya Weighted Edit Distance	0.610
VA Bhattacharyya Divergence	0.631
Bhattacharyya Total Weighted Edit Distance	0.646

Table 1. Squared correlation scores between the various model-based divergence measures and the empirical word confusabilities $-\log \mathrm{E}(w_1, w_2)$. VA refers to the variational HMM approximation of KL divergence or Bhattacharyya divergence. Min and Resistor are the two symmetrization methods. MC 100K refers to Monte Carlo simulations with 100,000 samples of HMM sequences.

Figure 3 shows a scatter plot of the variational KL divergence 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 also computed the HMM KL divergence by direct Monte-Carlo sampling of the HMM state sequences, as well as the Bhattacharyya and weighted edit distance methods. The variational bound was excluded because it did not perform as well as the variational approximation. Table 1 shows the results using all the different methods. The variational HMM KL divergence is about as good as the more accurate and time-consuming Monte Carlo estimates of KL divergence. Unfortunately, the HMM KL divergence itself is apparently not as well suited to the confusability task as the weighted edit distances and the Bhattacharyya divergence. This is natural since the Bhattacharyya divergence is known to yield a tighter bound on the Bayes error than the KL divergence. It is a bit surprising, though, that the Bhattacharyya total weighted edit distance outperforms the variational Bhattacharyya divergence, since it is actually a looser bound on the Bayes error. However, the confusability measurements produced by the recognizer are only loosely related to the Bayes error, because for example, the recognizer computes the Viterbi path instead of summing over paths.

8. REFERENCES

[1] Solomon Kullback, *Information Theory and Statistics*, Dover Publications Inc., Mineola, New York, 1968.

[2] Peder Olsen and Satya Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models," in *Eurospeech*, Geneva, Switzerland, September 1-4 2003, vol. 4, pp. 2509–2512.

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

[4] Jorge Silva and Shrikanth 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.

[5] 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 *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, PA, 2006, pp. 2338–2341.

[6] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," in *Proceedings of ICCV 2003*, Nice, October 2003, vol. 1, pp. 487–493.

[7] L. D. Brown, *Fundamentals of statistical exponential families. vol 9 of Lecture Notes - Monograph Series*, Institute of Math. Stat., 1991.

[8] Peder A. Olsen and Karthik Visweswariah, "Fast clustering of gaussians and the virtue of representing gaussians in exponential model format," *Proceedings of the International Conference on Spoken Language Processing*, October 2004.

[9] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T technical Journal*, vol. 64, no. 2, pp. 391–408, 1985.

[10] D. Johnson and S. Sinanovic, "Symmetrizing the Kullback Leibler distance," `http://www-dsp.rice.edu/~dhj/resistor.pdf`.

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

[12] Jorge Silva and Shrikanth Narayanan, "Upper bound Kullback-Leibler divergence for hidden Markov models with application as discrimination measure for speech recognition," *IEEE International Symposium on Information Theory*, 2006.

[13] Peder Olsen and John Hershey, "Bhattacharyya error and divergence using variational importance sampling," in *ICSLP*, Antwerp, Belgium, August 2007.