EXTENDED BAUM-WELCH REESTIMATION OF GAUSSIAN MIXTURE MODELS BASED ON REVERSE JENSEN INEQUALITY

Mohamed Afify

BBN Technologies, Cambridge, MA, 02138, USA

ABSTRACT

In this paper we derive the well known EBW reestimation formulae for Gaussian mixture models using the recently proposed reverse Jensen inequality. In addition to the simplicity of the derivation, it leads to closed form expressions for the D of each Gaussian in the mixture. Using some approximations, it is shown that the expressions can be reduced to the popular formula of [8] with a Gaussian dependent step size. The average of the normalized distance used in calculating the step size is empirically verified to be very close to one. Hence, the proposed formula leads to a global value very close to 3. The obtained results are validated in experiments on large vocabulary speech recognition.

1. INTRODUCTION

Discriminative training methods like maximum mutual information (MMI) and minimum phone error (MPE) estimation have recently witnessed increased interest from the speech recognition community for building large scale hidden Markov models (HMMs). These algorithms rely on using lattices for collecting the relevant statistics, and the so called extended Baum-Welch (EBW) reestimation formulae for estimating the means and variances of the models' distributions. One of the contributions of [8] is providing a recipe for calculating certain constants (called D) which are needed in the EBW equations, and their proper choice is crucial for the success of the training.

EBW reestimation was introduced for discrete distributions in [3], and first extended to Gaussian distributions based on a discrete approximation in [7]. This was followed by two separate proofs for the Gaussian case in [4], and [6]. As far as the D constants are concerned all these proofs are existence proofs, i.e. they show that there exists a large enough constant D that ensures the convergence of the estimates, without specifying the value of this constant.

In this paper EBW estimation of Gaussian mixture models is derived from a straightforward application of the reverse Jensen inequality, and the associated conditional expectation maximization (CEM) algorithm introduced in [5]. The case of discrete (multinomial) distributions also follows by applying the same framework, but is not considered here. The derivation first bridges the gap between these seemingly different methods for the optimization of discriminant objective functions, and offers an interesting alternative view to EBW estimation. In addition, it results in an analytical expression for the D constant for each Gaussian distribution in the mixture. From this point of view, and in contrast to proofs existing in the literature, the proof is constructive. The resulting expression for the D bears interesting relationships, under some approximations, to the value presented in [8].

The paper is organized as follows. First the basic principle of the proposed derivation is given in Section 2, this is followed by briefly reviewing the reverse Jensen inequality, and the CEM algorithm from [5] in Section 3. Section 4 derives EBW reestimation for Gaussian mixtures using the reverse Jensen inequality. It also gives an analytical expression for the D of each Gaussian component of the mixture which ensures the convergence of the estimates. Some approximations, and relationships to the D value introduced in [8] are discussed in Section 5. Section 6 verifies the obtained results in large vocabulary speech recognition experiments. We finally summarize our findings in Section 7.

2. BASIC PRINCIPLE

This section outlines the basic principle of optimizing discriminant objective functions for mixture models using the reverse Jensen inequality. To be concrete, we consider maximizing the MMI objective function, which can be written as

$$F(\theta) = \sum_t log P_\theta(o_t|\mathcal{M}^{num}) - \sum_t log P_\theta(o_t|\mathcal{M}^{den}) \quad (1)$$

where t, and m, vary over the observation and mixture indices respectively, o_t stands for the observation at time t, θ corresponds to the model parameters, and $\mathcal{M}^{num}$, and $\mathcal{M}^{den}$ are the numerator and denominator models respectively. We are interested in optimizing this objective function for mixture models, where

$$P_\theta(o_t|\mathcal{M}) = \sum_m P(m) P_\theta(o_t|m, \mathcal{M}) \quad (2)$$

and $P(m)$ are the mixture weights, and $P_\theta(o_t|m, \mathcal{M})$ are the mixture component distributions. We omit the dependence of the mixture weights on the parameter θ to highlight our interest in only estimating the component distribution parameters.

The first term on the right hand side of Equation (1) (denote it $L^{num}(\theta)$) is similar to the ML case. An auxiliary function ($Q^{num}(\theta, \hat{\theta})$), which is a lower bound to the likelihood, can be formed and iteratively optimized. This function is derived using the well known Jensen inequality [2], and can be written as

$$\begin{aligned} L^{num}(\theta) &\geq Q(\theta, \hat{\theta}) + C \\ &= \sum_t \sum_m \gamma_{m,t}^{num} \log P_\theta(o_t|m, \mathcal{M}^{num}) + C \end{aligned} \quad (3)$$

where $\hat{\theta}$ are the model parameters from the previous iteration, $\gamma_{m,t}^{num}$ are the component posteriors at time t computed using the previous model parameters, and C is a constant that absorbs all quantaties that do not depend on the current model parameters,

and hence are not important for optimization. It is known that iteratively maximizing the auxiliary function in Equation (3) leads to a monotonic increase of the log likelihood.

Inorder to lower bound the whole objective function, as in the ML case, we need to upper bound (instead of lower bound) the denominator term (denote it $L^{den}(\theta)$). If this bound (denote it $S^{den}(\theta,\hat{\theta})$) exists we can directly construct an objective function $R(\theta,\hat{\theta}) = Q^{num}(\theta,\hat{\theta}) - S^{den}(\theta,\hat{\theta})$ which can be iteratively optimized similar to the ML case, and which ensures the increase of the original objective function in Equation (1). Moreover, it would be interesting, from the point of view of ease of optimization, if we have an upper bound to the denominator likelihood that has the same form as Equation (3), i.e. a weighted sum of complete data log likelihoods. This can be written as

$$\begin{aligned} L^{den}(\theta) &\leq S^{den}(\theta,\hat{\theta}) + B \\ &= \sum_t \sum_m (-w_{m,t}) \log P_\theta(z_{m,t}|m,\mathcal{M}) + B \end{aligned} \quad (4)$$

where B is a constant depending only on the previous model parameters, $w_{m,t}$, and $z_{m,t}$ are positive weights, and modified observations that will be specified later.

The bound in Equation (4) clearly looks like a reverse Jensen inequality, and the associated optimization technique of discriminant objective functions is referred to as the conditional expectation-maximization (CEM) algorithm. Both were introduced in [5], and will be overviewed in the following section.

3. REVERSE JENSEN AND THE CEM ALGORITHM

The existence of a reverse Jensen bound for the sake of constructing an auxiliary function for discriminative criteria was addressed in [5]. It was shown that such bound can be found for mixtures of exponential family members, and provided explicit solutions for the weights and the modified observations like those in Equation (4). Specifically, it was shown that

$$\begin{aligned} \log p_\theta(X) &= \log \sum_m P(m) p_\theta(X|m) \\ &= \log \sum_m P(m) \exp(A(X) + \theta_m^T X - K(\theta_m)) \\ &\leq \log \sum_m (-w_m) \log p_\theta(Y_m|m) + B \end{aligned} \quad (5)$$

where the second line in Equation (5) is a parametrization of the exponential family of distributions, and the third line is the reverse Jensen bound with the B term absorbing all constants. The positive weights w_m, and the component dependent observations Y_m are given by

$$Y_m = \frac{\gamma_m}{w_m}\left(\frac{\partial K(\theta_m)}{\partial \theta_m}\Big|_{\hat{\theta}_m} - X\right) + \frac{\partial K(\theta_m)}{\partial \theta_m}\Big|_{\hat{\theta}_m} \quad (6)$$

$$\begin{aligned} w_m &= 4G(\gamma_m/2)(X - K'(\hat{\theta}_m))^T K''(\hat{\theta}_m)^{-1}(X - K'(\hat{\theta}_m)) \\ &\quad + w'_m \end{aligned} \quad (7)$$

$$\begin{aligned} w'_m &= \min w'_m \quad \text{s.t.} \\ &\frac{\gamma_m}{w'_m}\left(\frac{\partial K(\theta_m)}{\partial \theta_m}\Big|_{\hat{\theta}_m} - X\right) + \frac{\partial K(\theta_m)}{\partial \theta_m}\Big|_{\hat{\theta}_m} \in \frac{\partial K(\theta_m)}{\partial \theta_m} \end{aligned} \quad (8)$$

where Y_m is the transformed observation, γ_m is the usual posterior for mixture component m, and the derivatives are calculated at the old parameter values $\hat{\theta}_m$. $K'()$, and $K''()$ are shorthand for the gradient and the Hessian of $K()$, and $()^T$, and $()^{-1}$ stand for transpose, and matrix inverse respectively. The w'_m is the smallest value of the weight that ensures that the transformed observations are in the gradient space of $K()$. This is similar in principle to the usual condition of maintaining positive variances during discriminative optimization. Finally the $G()$ is a nonlinear function that is obtained by trying to maintain a tight bound. For further details we refer the reader to [5].

By applying both Jensen and reverse Jensen to bound both the numerator, and denominator terms, and differentiating the resulting function and equating the derivative to zero. We arrive at the following solution for the parameters of the m^{th} mixture component. This is referred to as the CEM algorithm.

$$\frac{\partial K(\theta_m)}{\partial \theta_m}\sum_t \gamma_{m,t}^{num} + w_{m,t}^{den} = \sum_t \gamma_{m,t}^{num} X_t + \sum_t w_{m,t}^{den} Y_{m,t} \quad (9)$$

where summation is over observations (indexed by t), and we have augmented the posteriors and weights by numerator (num), and denominator (den) labels.

4. EBW REESTIMATION FOR GAUSSIAN MIXTURES

In this section we will show that a straightforward application of the above formulation in the Gaussian mixture case will lead us to the well known EBW reestimation formulae. In addition, an explicit formula for the D for each Gaussian component is provided. We will focus on single dimension observations which can be readily extended to the vector case. First rewriting a Gaussian distribution $p(o|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(o-\mu)^2}{2\sigma^2}\right)$ in the exponential family notation we can observe that

$$X = (o \quad \frac{-1}{2}o^2)^T \quad (10a)$$

$$\theta = (\frac{\mu}{\sigma^2} \quad \frac{1}{\sigma^2})^T \equiv (\theta_1 \quad \theta_2)^T \quad (10b)$$

$$K(\theta) = \frac{1}{2}(\frac{\mu^2}{\sigma^2} - \log(\frac{1}{\sigma^2})) \equiv \frac{1}{2}(\frac{\theta_1^2}{\theta_2} - \log(\theta_2)) \quad (10c)$$

$$A(X) = \frac{-1}{2}\log(2\pi) \quad (10d)$$

and the gradient, and the inverse Hessian of $K(\theta)$ can be also derived as

$$K'(\theta) = (\frac{\theta_1}{\theta_2} \quad \frac{-\theta_1^2}{2\theta_2^2} - \frac{1}{2\theta_2})^T \quad (11a)$$

$$K''(\theta)^{-1} = \begin{pmatrix} 2\theta_1^2 + \theta_2 & 2\theta_1\theta_2 \\ 2\theta_1\theta_2 & 2\theta_2^2 \end{pmatrix} \quad (11b)$$

where we note that the dependence on the mixture component m has been dropped for convenience in these definitions.

Now, by using the exponential family form of the Gaussian distribution and the gradient and Hessian in Equations (10), and (11). Then by applying the definitions of Y_m, and w_m from Equations (6), and (7) in Equation (9), and solving for μ_m, and σ_m^2. We arrive at the well known EBW reestimation formulae for the

mean and variance of each mixture component, which we rewrite for reference as

$$\mu_m = \frac{\sum_t \gamma_{m,t}^{num} o_t - \sum_t \gamma_{m,t}^{den} o_t + D_m \hat{\mu}_m}{\sum_t \gamma_{m,t}^{num} - \sum_t \gamma_{m,t}^{den} + D_m} \quad (12)$$

$$\sigma_m^2 = \frac{\sum_t \gamma_{m,t}^{num} o_t^2 - \sum_t \gamma_{m,t}^{den} o_t^2 + D_m(\hat{\mu}_m^2 + \hat{\sigma}_m^2)}{\sum_t \gamma_{m,t}^{num} - \sum_t \gamma_{m,t}^{den} + D_m} - \mu_m^2 \quad (13)$$

where D_m is given by

$$D_m = \sum_t \gamma_{m,t}^{den} + \sum_t w_{m,t}^{den} \equiv \gamma_m^{den} + w_m^{den} \quad (14)$$

and from Equation (7) w_m^{den} is evaluated as

$$\begin{aligned} w_m^{den} &= \sum_t w_{m,t}^{\prime den} + 4 \sum_t G(\gamma_{m,t}^{den}/2)(X_t - K'(\hat{\theta}_m))^T \times \\ &\quad K''(\hat{\theta}_m)^{-1}(X_t - K'(\hat{\theta}_m)) \\ &\equiv w_m^{\prime den} + w_m^{\prime\prime den} \end{aligned} \quad (15)$$

where from the condition in Equation (8) we have

$$\begin{aligned} w_m^{\prime den} &= \sum_t \max\left(\gamma_{m,t}^{den}\left(\frac{o_t^2}{\hat{\mu}_m^2 + \hat{\sigma}_m^2} - 1\right), 0\right) \\ &= \sum_{t: o_t^2 > \hat{\mu}_m^2 + \hat{\sigma}_m^2} \gamma_{m,t}^{den}\left(\frac{o_t^2}{\hat{\mu}_m^2 + \hat{\sigma}_m^2} - 1\right) \end{aligned} \quad (16)$$

and after some algebraic manipulations we can simplify $w_m^{\prime\prime den}$ to

$$\begin{aligned} w_m^{\prime\prime den} &= 4 \sum_t G(\gamma_{m,t}^{den}/2) \frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} \\ &\quad + 4 \sum_t G(\gamma_{m,t}^{den}/2)\left(\frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} - 1\right)^2 \\ &\equiv w_{m,1}^{\prime\prime den} + w_{m,2}^{\prime\prime den} \end{aligned} \quad (17)$$

Now using Equations (14)-(17) we write a closed form expression for the D_m of each Gaussian component as

$$D_m = \gamma_m^{den} + w_m^{\prime den} + w_{m,1}^{\prime\prime den} + w_{m,2}^{\prime\prime den} \quad (18)$$

which guarantees the increase of the discriminative objective function at each iteration. In the following section, we will give two approximations which will allow calculating the D constants from the statistics of the mean and the variance.

5. APPROXIMATIONS AND INSIGHTS

This section first proposes two approximations that will allow us to write the values of $w_{m,1}^{\prime\prime den}$, and $w_{m,2}^{\prime\prime den}$ from the mean and variance statistics. These are given as

$$\begin{aligned} w_{m,1}^{\prime\prime den} &= 4 \sum_t G(\gamma_{m,t}^{den}/2) \frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} \\ &\approx 2 \sum_t \gamma_{m,t}^{den} \frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} \\ &= 2\gamma_m^{den}\left[\frac{(\mu_m^{den} - \hat{\mu}_m)^2 + \sigma_m^{den^2}}{\hat{\sigma}_m^2}\right] \end{aligned} \quad (19)$$

where the second line follows from the approximation $G(a) \approx a$ which is accurate for $a \geq 1/6$ [5], and the third line is obtained by straightforward algebraic simplification. Similarly

$$\begin{aligned} w_{m,2}^{\prime\prime den} &= 4 \sum_t G(\gamma_{m,t}^{den}/2)\left(\frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} - 1\right)^2 \\ &\approx 2 \sum_t \gamma_{m,t}^{den}\left(\frac{(o_t - \hat{\mu}_m)^2}{\hat{\sigma}_m^2} - 1\right)^2 \\ &\approx 2\gamma_m^{den}\left(\frac{(\mu_m^{den} - \hat{\mu}_m)^2 + \sigma_m^{den^2}}{\hat{\sigma}_m^2} - 1\right)^2 \end{aligned} \quad (20)$$

where the second line follows from the approximation $G(a) \approx a$, and the third line from the approximation $E[f^2(x)] \approx E^2[f(x)]$. In both equations we have μ_m^{den}, and $\sigma_m^{den^2}$ are similar to maximum likelihood estimates but calculated using the denominator posteriors. Using these approximations we can rewrite the D_m in Equation (18) as

$$D_m \approx \gamma_m^{den}[1 + 2Z_m] + w_m^{\prime den} \quad (21)$$

where $Z_m = U_m + (U_m - 1)^2$, and U_m is a normalized distance given by

$$U_m = \left[\frac{(\mu_m^{den} - \hat{\mu}_m)^2 + \sigma_m^{den^2}}{\hat{\sigma}_m^2}\right] \quad (22)$$

Now recall that [8] calculates $D_m = \max(D'_m, E\gamma_m^{den})$, where D'_m is calculated to ensure that the variance is positive[1], and E is globally chosen in the interval $[1.0 - 2.0]$ which ensures a reasonable convergence behavior. Note that $w_m^{\prime den}$ comes from a similar condition to the positivity of the variance as discussed in Section 3. So if we replace $w_m^{\prime den}$ by D'_m, and the sum by taking the maximum in Equation (21), we can finally write $D_m = \max(D'_m, E_m\gamma_m^{den})$, where E_m is shorthand for the term multiplying the denominator posterior. This is exactly like the formula in [8] but with the global constant E replaced by a Gaussian dependent constant E_m. The latter is proportional to the normalized distance U_m of Equation (22).

6. EXPERIMENTAL EVALUATION

The proposed formula was tested in Arabic broadcast news (BN) transcription experiments. The acoustic models are built, from about 100 hours of Arabic BN data. The parameter space is of dimension 46. This is obtained by HLDA transforming a 60-dimension feature vector composed of 14 cepstra, energy, and their first, second, and third derivatives. Recognition consists of an unadapted decoding pass which provides superviosion to a second adapted decoding pass. MMI estimation is limited to the SAT models used in adapted decoding. This model is a non-crossword state clustered tied mixture (SCTM_NX) model, that has about 150K Gaussain distributions with diagonal covariances. six iterations of MMI estimation are used to refine the original ML SAT model. The ML model is also used to construct lattices required in MMI estimation. The language model factor used in MMI training is set to 15. This was tuned in earlier experiments to give the best performance on this task. Results are reported on the dev04 test set. Similar findings were also observed on dev03 set, but are not shown here.

[1] Usually the minimum value is multiplied by 2 in practice.

The ML trained system has an error rate of 18.8%. More details about the system structure can be found in [1].

In using the proposed formula, the value of Z_m, in Equation (21), is averaged over the 46 vector dimensions to come up with a single estimate for each Gaussian distribution. The mean and variance of Z_m, for all Gaussian distributions in the above model, are shown in Table 1 for each iteration.

Mean	Variance
0.999108	0.001441
1.001699	0.001463
1.004833	0.001574
1.008308	0.001844
1.011828	0.001970
1.015538	0.002477

Table 1: The mean and the variance of Z_m calculated over 150k Gaussians for 6 iterations of MMI training of a non-crossword acoustic model.

As can be observed from the table, the mean is always very close to one with a very small variance. This in turn implies that the mean value of the normalized distance U_m, in Equation (22), is also very close to one. A possible explanation is that this normalized distance U_m, as can be observed from the second line of Equation (19), is basically a ratio of two variances that although calculated using different statistics (the MMI vs denominator) turn out to be very close in practice. On the otherhand, an extreme case that leads to $Z_m \approx 1$ is as follows. Take the first iteration, where the model is initialized from the ML (equivalently numerator) estimates, as example. A value of $Z_m \approx 1$ may indicate that $\mu_m^{num} \approx \mu_m^{den}$, and $\sigma_m^{num^2} \approx \sigma_m^{den^2}$. This might lead us to think that the denimenator lattice is not deep enough, or an improper LM factor is used. This did not turn out to be the case in our experiments. We are currently verifying these hypotheses in English CTS experiments.

In any case, for a value of the normalized distance that is close to one, the proposed formula will be very similar to a global step size equal to three. To further verify this, we compare the recognition accuracy for the proposed formula, and the constant E result for both $E = 1$, and $E = 2$. These are shown in Table 2. In the results the 2 multiplicative factor in Equation (21) is ignored, and hence the proposed formula should be comparable to the case with constant $E = 2$. This can be seen from the table, where both the proposed formula, and the constant $E = 2$ have equal WER. They also, although not shown here, have almost identical MMI scores during the training. The value $E = 1$ has also similar WER, although it leads to faster convergence and hence a better MMI score at the end of the training.

Roughly speaking, if the result $Z_m \approx 1$ is valid, the developments in this paper may be a theoretical justification for using a constant multiple of the denominator count as the D value as suggested in [8]. If the approximations made in Section 5 still maintain the theoretical guarantee then a value of $E = 3$ still ensures the increase of the objective function. While in practice smaller values of $E = 1$, or $E = 2$ work well, their verification might need the development of tighter bounds.

E formula	WER
$E = 1$	18.4
$E = 2$	18.4
Proposed	18.4

Table 2: The WER for adapted decoding on the dev04 set for Arabic BN transcription using the proposed E formula and global $E = 1$, and $E = 2$.

7. SUMMARY

In this paper we derived the well known EBW reestimation formulae for Gaussian mixture models using the recently proposed reverse Jensen inequality. In contrast to earlier derivations we provide closed form solution for the smoothing constant D of each Gaussian. Some approximations reduce the D formula to an expression very close to the popular proposal of [8], but that provides a Gaussian dependent step size instead of the global value known as E. In speech recognition experiments we verify that the normalized distance used in the calculation is very close to one for all Gaussian distributions, and hence leads a global value of $E \approx 3$.

8. ACKNOWLEDGEMENT

The authors would like to thank Thomas Colthurst, Spyros Matsoukas, and Bing Xiang, from BBN Technologies for reading initial versions of the paper, and for useful discussions.

9. REFERENCES

[1] M. Afify, et al., "The BBN Non-English evaluation systems for broadcast news," in Proc. of the RT04 workshop, Palisades, NY, November 2004.

[2] T. Cover and J. Thomas, Elements of Information Theory, Wiley & Sons, New York, 1991.

[3] P.S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," IEEE Trans on Information Theory, vol. 37, no. 1, pp. 107-113, Jan. 1991.

[4] A. Gunawardana, and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in Proceedings Eurospeech'01, Aalborg, Denmark, September 2001.

[5] T. Jebara, Discriminative, Generative, and Imitative Learning, Ph.D. thesis Massachusetts Institute of Technology, February 2002.

[6] D. Kanevsky, "Extended Baum transformations for general functions," in Proceedings ICASSP'04, Montreal, Canada, May 2004.

[7] Y. Normandin, R. Cardin, and R. De Mori, "High-performance connected digit recognition using maximum mutual information estimation," IEEE Trans on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994.

[8] P.C. Woodland, and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," Computer Speech and Language, Vol. 16, pp. 25-48, 2002.