EXTENDED BAUM TRANSFORMATIONS FOR GENERAL FUNCTIONS

*Dimitri Kanevsky**

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
{ kanevsky@ us.ibm.com }

ABSTRACT

The discrimination technique for estimating the parameters of Gaussian mixtures that is based on the Extended Baum transformations (EB) has had significant impact on the speech recognition community. There appear to be no published proofs that definitively show that these transformations increase the value of an objective function with iteration (i.e., so-called "growth transformations"). The proof presented in the current paper is based on the linearization process and the explicit growth estimate for linear forms of Gaussian mixtures. We also derive new transformation formulae for estimating the parameters of Gaussian mixtures generalizing the EB algorithm, and run simulation experiments comparing different growth transformations.

1. INTRODUCTION

The EB procedure involves two types of transformations that can be described as follows. Let $F(z) = F(z_{ij})$ be some function in variables $z = (z_{ij})$ and $c_{ij} = z_{ij}\frac{\delta}{\delta z_{ij}}F(z)$.
I. Discrete probabilities:

$$\hat{z}_{ij} = \frac{c_{ij} + z_{ij}C}{\sum_i c_{ij} + C} \tag{1}$$

where $z \in D = \{z_{ij} \geq 0, \sum_j z_{ij} = \sum_{j=1}^{j=m_i} z_{ij} = 1\}$

II. Gaussian mixture densities:

$$\hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i\in I} c_{ij}y_i + C\mu_j}{\sum_{i\in I} c_{ij} + C} \tag{2}$$

$$\hat{\sigma}_j^2 = \hat{\sigma}_j(C)^2 = \frac{\sum_{i\in I} c_{ij}y_i^2 + C(\mu_j^2 + {\sigma_j}^2)}{\sum_{i\in I} c_{ij} + C} - \hat{\mu}_j^2 \tag{3}$$

where

$$z_{ij} = \frac{1}{(2\pi)^{1/2}\sigma_j}e^{-(y_i-\mu_j)^2/2\sigma_j^2} \tag{4}$$

and y_i is a sample of training data. It was shown in [4] that (1) are growth transformations for sufficiently large C when F is a rational function. Updated formulae (2, 3) for rational functions F were obtained through discrete probability approximation of Gaussian densities [6] and have been widely used as an alternative to direct gradient-based optimization approaches ([8], [7]). As originally presented in our IBM Research Report [5] we demonstrate in this paper that (1) and (2, 3) are growth transformations for sufficiently large C if functions F obey certain smoothness constraints. Axelrod [1] has recently proposed another proof of existence of a constant C that ensures validity of the MMIE auxiliary function as formulated by Gunawardana et al. [3]).

*The work is partly supported by the DARPA Babylon Speech-to-Speech Translation Program.

2. LINEARIZATION

This principle is needed to reduce proofs of growth transformation for general functions to linear forms.

Lemma 1 *Let*

$$F(z) = \tilde{F}(\{u_j\}) = \tilde{F}(\{g_j(z)\}) = \tilde{F} \circ g(z) \tag{5}$$

where $u_j = g_j(z), j = 1, ..m$ and z varies in some real vector space R^n of dimension n. Let $g_j(z)$ for all $j = 1, ...m$ and $F(z)$ be differentiable at z. Let, also, $\frac{\delta\tilde{F}(\{u_j\})}{\delta u_j}$ exist at $u_j = g_j(z)$ for all $j = 1, ...m$. Let, further, $L(z\prime) \equiv \nabla\tilde{F}\Big|_{g(z)} \cdot g(z\prime)$, $z\prime \in R^n$. Let T_C be a family of transformations $R^n \to R^n$ such that for some $l = (l_1...l_n) \in R^n$ $T_C(z) - z = l/C + o(1/C)$ if $C \to \infty$. (Here $o(\epsilon)$ means that $o(\epsilon)/\epsilon \to 0$ if $\epsilon \to 0$). Let, further, $T_C(z) = z$ if

$$\nabla L|_z \cdot l = 0 \tag{6}$$

Then for sufficiently large C T_C is growth for F at z iff T_C is growth for L at z.

Proof First, from the definition of L we have
$\frac{\delta F(z)}{\delta z_k} = \sum_j \frac{\delta\tilde{F}(\{u_j\})}{\delta u_j}\frac{\delta g_j(z)}{\delta z_k} = \frac{\delta L(z)}{\delta z_k}$
Next, for $z\prime = T_C(z)$ and sufficiently large C we have: $F(z\prime) - F(z) = \sum_i \frac{\delta F(z)}{\delta z_i}(z_i\prime - z_i) + o(1/C) = \sum_i \frac{\delta F(z)}{\delta z_i}l_i/C + o(1/C) = \sum_i \frac{\delta L(z)}{\delta z_i}l_i/C + o(1/C) = \sum_i \frac{\delta L(z)}{\delta z_i}(z_i\prime - z_i) + o(1/C) = L(z\prime) - L(z) + o(1/C)$. Therefore for sufficiently large C $F(z\prime) - F(z) > 0$ iff $L(z\prime) - L(z) > 0$.

3. EB FOR DISCRETE PROBABILITIES

The following theorem is a generalization of [4].

Theorem 1 *Let $F(z)$ be a function that is defined over $D = \{z_{ij} \geq 0, \sum z_{ij} = 1\}$. Let F be differentiable at $z \in D$ and let $\hat{z} \neq z$ be defined as in (1). Then $F(\hat{z}) > F(z)$ for sufficiently large positive C and $F(\hat{z}) < F(z)$ for sufficiently small negative C.*

Proof Following the linearization principle, we first assume that $F(z) = l(z) = \sum a_{ij} z_{ij}$ is a linear form. Than the transformation formula for $l(x)$ is the following:

$$\hat{z}_{ij} = \frac{a_{ij} z_{ij} + C z_{ij}}{l(z) + C} \tag{7}$$

We need to show that $l(\hat{z}) \geq l(z)$. It is sufficient to prove this inequality for each linear sub component associated with i

$$\sum_{j=1}^{j=n} a_{ij} \hat{z}_{ij} \geq \sum_{j=1}^{j=n} a_{ij} z_{ij}$$

Therefore without loss of generality we can assume that i is fixed and drop subscript i in the forthcoming proof (i.e. we assume that $l(z) = \sum a_j z_j$, where $z = \{z_j\}$, $z_j \geq 0$ and $\sum z_j = 1$). We have: $l(\hat{z}) = \frac{l_2(z) + C l(z)}{l(z) + C}$, where $l_2(z) := \sum_j a_j^2 z_j$. The linear case of Theorem 1 will follow from next two lemmas.

Lemma 2

$$l_2(z) \geq l(z)^2 \tag{8}$$

Proof Let as assume that $a_j \geq a_{j+1}$ and substituting $z\prime = \sum_{j=1}^{j=n-1} z_j$ we need to prove:

$$\sum_{j=1}^{j=n-1} [a_j^2 z_j + a_n^2 (1 - z\prime)] \geq \sum_{j=1}^{j=n-1} (a_j - a_n)^2 z_j^2 + 2 \sum_{j=1}^{j=n-1} (a_j - a_n) a_n z_j + a_n^2 \tag{9}$$

We will prove the above formula by proving for every fixed j $(a_j^2 - a_n^2) z_j \geq (a_j - a_n)^2 z_j^2 + 2(a_j - a_n) a_n z_j$. If $(a_j - a_n) z_j \neq 0$ then the above inequality is equivalent to $a_j - a_n \geq (a_j - a_n) z_j$ and is obviously holds since $0 \leq z_j \leq 1$

Lemma 3 *For sufficiently large $|C|$ the following holds:*

$l(\hat{z}) > l(z)$ if C is positive and $l(\hat{z}) < l(z)$ if C is negative.

Proof From (8) we have the following inequalities.
$l_2(z) + C l(z) \geq l(z)^2 + C l(z)$,
$l(\hat{z}) = \frac{l_2(z) + C l(z)}{l(z) + C} \geq \frac{l(z)^2 + C l(z)}{l(z) + C}$ if $l(z) + C > 0$
and $l(\hat{z}) = \frac{l_2(z) + C l(z)}{l(z) + C} \leq \frac{l(z)^2 + C l(z)}{l(z) + C}$ if $l(z) + C < 0$.
The general case of Theorem 1 follows immediately from the observation that (6) is equivalent to $l_2(z) - l(z)^2 = 0$ for large C.

4. EB FOR GAUSSIAN DENSITIES

For simplicity of the notation we consider the transformation (2), (3), only for a single pair of variables μ, σ, i.e. we drop subscript j everywhere in (2, 3), (4) and also set $\hat{z}_i = \frac{1}{(2\pi)^{1/2} \hat{\sigma}} e^{-(y_i - \hat{\mu})^2 / 2\hat{\sigma}^2}$

Theorem 2 *Let $F(\{z_i\})$, $i = 1...m$, be differentiable at μ, σ and $\frac{\delta F(\{z_i\})}{\delta z_i}$ exist at z_i. Let either $\hat{\mu} \neq \mu$ or $\hat{\sigma} \neq \sigma$. Then for sufficiently large C*

$$F(\{\hat{z}_i\}) - F(\{z_i\}) = T/C + o(1/C) \tag{10}$$

Where

$$T = \frac{1}{\sigma^2} \{ \frac{\{\sum c_j [(y_j - \mu)^2 - \sigma^2]\}^2}{2\sigma^2} + [\sum c_j (y_j - \mu)]^2 \} > 0 \tag{11}$$

In other words, $F(\{\hat{z}_i\})$ grows proportionally to $1/C$ for sufficiently large C.

Proof First, we assume that $F(\{z_i\}) = l(\mu, \sigma) := l(\{z_i\}) := \sum_{i=1}^{i=m} a_i z_i$. Let us set $l(\hat{\mu}, \hat{\sigma}) := l(\{\hat{z}_i\}) := \sum_{i=1}^{i=m} a_i \hat{z}_i$. Then $c_j = a_j z_j$ in (2), (3). We want to prove that for sufficiently large C $l(\hat{\mu}, \hat{\sigma}) \geq l(\mu, \sigma)$. This inequality is sufficiently to prove with the precision $1/C^2$.

$$\hat{\mu} = \hat{\mu}(C) = \frac{\sum_{\mathrm{J}=1}^{j=m} c_j y_j + C\mu}{\sum_{\mathrm{J}=1}^{j=m} c_j + C} = \frac{\frac{1}{C} \sum_{j=1}^{j=m} c_j y_j + \mu}{\frac{1}{C} \sum_{\mathrm{J}=1}^{j=m} c_j + 1} \sim$$

$$\sim (\frac{1}{C} \sum_j c_j y_j + \mu)(1 - \frac{\sum_j c_j}{C}) \sim \mu + \frac{1}{C} (\sum_j c_j y_j - \mu \sum_j c_j) \tag{12}$$

$$\hat{\mu} \sim \mu + \frac{\sum_j [c_j (y_j - \mu)]}{C} \tag{13}$$

Next, we have

$$\hat{\sigma}^2 = \hat{\sigma}(C)^2 = \frac{\sum_j c_j y_j^2 + C(\mu^2 + \sigma^2)}{\sum_j c_j + C} - \hat{\mu}^2 \tag{14}$$

Let us compute $\hat{\sigma}^2$ using (14)

$$\frac{\sum_j c_j y_j^2 + C(\mu^2 + \sigma^2)}{\sum_j c_j + C} \sim$$

$$\sim (\frac{\sum_j c_j y_j^2}{C} + \mu^2 + \sigma^2)(1 - \frac{\sum_j c_j}{C}) \sim$$

$$\sim \mu^2 + \sigma^2 + \frac{1}{C} [\sum_j c_j y_j^2 - (\mu^2 + \sigma^2) \sum_j c_j] \tag{15}$$

$$\hat{\mu}^2 \sim \mu^2 + \frac{2\mu}{C} \sum_{\mathrm{J}=1}^{j=m} c_j (y_j - \mu) \tag{16}$$

This gives

$$\hat{\sigma}^2 \sim \mu^2 + \sigma^2 + \frac{1}{C}[\sum_j c_j y_j^2 - (\mu^2+\sigma^2)\sum_j c_j] -$$

$$-[\mu^2 + \frac{2\mu}{C}\sum_j c_j(y_j-\mu)] =$$

$$= \sigma^2 + \frac{1}{C}[\sum_j c_j y_j^2 - (\mu^2+\sigma^2)\sum_j c_j - 2\mu\sum_j c_j(y_j-\mu)] \quad (17)$$

And finally

$$\hat{\sigma}^2 \sim \sigma^2 + \frac{\sum_j[(y_j-\mu)^2-\sigma^2]c_j}{C} \quad (18)$$

$$(y_i-\hat{\mu})^2/\hat{\sigma}^2 \sim \frac{1}{\sigma^2}[(y_i-\mu)^2 -$$

$$-\frac{2(y_i-\mu)\sum_j c_j(y_j-\mu)}{C}]\times$$

$$\times\{1 - \frac{\sum_j c_j[(y_j-\mu)^2-\sigma^2]}{\sigma^2 C}\} \sim$$

$$\sim \frac{(y_i-\mu)^2}{\sigma^2} - \frac{1}{C\sigma^2}\{\frac{(y_i-\mu)^2}{\sigma^2}\sum_j[(y_j-\mu)^2-\sigma^2]c_j +$$

$$+2(y_i-\mu)\sum_j(y_j-\mu)c_j\} \quad (19)$$

$$\hat{z}_i \sim \frac{1}{(2\pi)^{1/2}\hat{\sigma}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}+\frac{A_i}{C\sigma^2}} \quad (20)$$

Where

$$A_i = \frac{(y_i-\mu)^2}{2\sigma^2}\sum_j[(y_j-\mu)^2-\sigma^2]c_j + (y_i-\mu)\sum_j(y_j-\mu)c_j$$

Continue this we have

$$\hat{z}_i \sim K e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}(1+\frac{A_i}{C\sigma^2}) \quad (21)$$

Where

$$K = \frac{1}{(2\pi)^{1/2}\hat{\sigma}}$$

$$1/\hat{\sigma} \sim \frac{1}{\sigma}\{1 - \frac{\sum_j c_j[(y_j-\mu)^2-\sigma^2]}{2\sigma^2 C}\} \quad (22)$$

$$(1+\frac{A_i}{C\sigma^2})\{1 - \frac{\sum_j c_j[(y_j-\mu)^2-\sigma^2]}{2\sigma^2 C}\} \sim$$

$$\sim 1 + \frac{1}{C\sigma^2}\{\frac{(y_i-\mu)^2}{2\sigma^2}\sum_j[(y_j-\mu)^2-\sigma^2]c_j +$$

$$+(y_i-\mu)\sum_j(y_j-\mu)c_j - 1/2\sum_j c_j[(y_j-\mu)^2-\sigma^2]\} \sim$$

$$\sim 1 + \frac{B_i}{C\sigma^2} \quad (23)$$

Where $B_i = [\frac{(y_i-\mu)^2}{2\sigma^2} - 1/2]\sum_j[(y_j-\mu)^2-\sigma^2]c_j + (y_i-\mu)\sum_j(y_j-\mu)c_j$

Using the last equalities we get

$$\hat{z}_i = z_i + \frac{B_i}{C\sigma^2}z_i \quad (24)$$

Since $l(\hat{\mu},\hat{\sigma})$ is a linear form in the z_i we have

$$l(\{\hat{z}_i\}) = l(\{z_i\}) + \frac{l(\{B_i z_i\})}{C\sigma^2} \quad (25)$$

and

$$l(\{B_i z_i\}) = \sum_i a_i z_i\{[\frac{(y_i-\mu)^2}{2\sigma^2} - 1/2]\times$$

$$\times\sum_j c_j[(y_j-\mu)^2-\sigma^2] + (y_i-\mu)\sum_j c_j(y_j-\mu)\} =$$

$$= \sum_i c_i\{[\frac{(y_i-\mu)^2}{2\sigma^2} - 1/2]\sum_j c_j[(y_j-\mu)^2-\sigma^2] +$$

$$+(y_i-\mu)\sum_j c_j(y_j-\mu)\} =$$

$$= \frac{\{\sum_j c_j[(y_j-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + [\sum_j c_j(y_j-\mu)]^2 \quad (26)$$

$$l(\{\hat{z}_i\}) - l(\{z_i\}) \sim \frac{T}{C}$$

Since by assumption either $\hat{\mu} \neq \mu$ or $\hat{\sigma} \neq \sigma$ $T \neq 0$. Applicability of the lineriazation principle follows from the fact that if (11) holds then the left part in the equation (6) is not equal to zero. Q.E.D.

5. NEW GROWTH TRANSFORMATIONS

One can derive new updates for means and variances applying EB algorithm of the section 3 by introducing probability constraints for means and variances as follows. Let us assume that $0 \leq \mu_j \leq D_j, 0 \leq \sigma_j \leq E_j$. Then we can introduce slack variables $\mu_j\prime \geq 0$, $\sigma_j\prime \geq 0$ such that $\mu_j/D_j + \mu_j\prime/D_j = 1$, $\sigma_j/E_j + \sigma_j\prime/E_j = 1$. Then we can compute updates as in (1), with c_j as in (2, 3).

$$\hat{\mu}_j = D_j\mu_j\frac{\sum_i c_{ij}\frac{(y_i-\mu_j)}{\sigma_j^2}+C}{\sum_i c_{ij}\frac{(y_i-\mu_j)}{\sigma_j^2}\mu_j + D_j C}$$

$$\hat{\sigma}_j = E_j\frac{\sum_i c_{ij}[-1+\frac{(y_i-\mu_j)^2}{\sigma_j^2}]+C\sigma_j}{\sum_i c_{ij}[-1+\frac{(y_i-\mu_j)^2}{\sigma_j^2}]+E_j C}$$

If some $\mu_j < 0$ one can make them positive by adding positive constants, compute updates for new variables in the new coordinate system and then go back to the old system of coordinates.

6. EXPERIMENTS AND DISCUSSION

Our preliminary experiments are done for a single pair of means and variances $0 < \mu < 3$, $0 < \sigma < 3$ (i.e. a subscript j can be dropped in update formulae in sections 1 and 5) and a Gaussian mixture $l(\mu, \sigma) = \sum_{i=1}^{i=100} a_i z_i$. Coefficients in this linear form a_i and y_i were chosen randomly. One iteration consists of three following steps:
1. *EB with the best C*: Compute $\mu_s = \hat{\mu}(C\prime)$, $\sigma_s = \hat{\sigma}(C\prime)$ as in (2, 3) where
$C\prime = argmax_{C\in\{1,2,\ldots,100\}} l(\hat{\mu}(C), \hat{\sigma}(C))$
2. *Modified EB with the best C:* Compute $\mu_m = \hat{\mu}(C\prime)$, $\sigma_m = \hat{\sigma}(C\prime)$ as in the section 5 where
$C\prime = argmax_{C\in\{1,\ldots,100\}} l(\hat{\mu}(C), \hat{\sigma}(C))$
3. *Mixture of EB and modified EB with the best C*: We define the best mixture as: $\mu(\tilde{\alpha}) = \alpha\mu_s + (1-\alpha)\mu_m$ and $\sigma(\tilde{\alpha}\prime) = \alpha\prime\sigma_s + (1-\alpha\prime)\sigma_m$ where
$(\tilde{\alpha}, \tilde{\alpha}\prime) = argmax_{(\alpha,\alpha\prime)\in[0,1]\times[0,1]} l(\mu(\alpha), \sigma(\alpha\prime))$
We repeatedly run three experiments (each consisting of 5 iterations: the EB (step 1), the modified EB (step 2) and the mixture (steps 1-3, in which an output from step 3 was fed as the input in the step 1, i.e. $(\mu, \sigma) = (\mu(\tilde{\alpha}), \tilde{\sigma}(\tilde{\alpha}\prime))$). A typical plot of three experiments is shown in Figure 1 (values of the objective function are placed along the ordinate axis). These illustrative simple numerical experiments

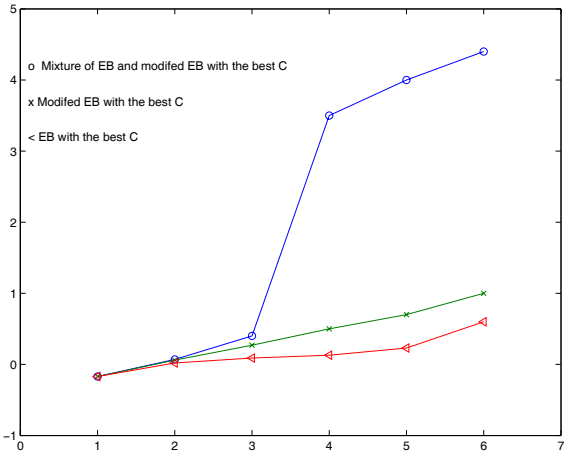

Fig. 1. Graphs of objective values for 3 maximization methods.

show that different growth transformations can exhibit different behavior and that combining them with appropriate weights can improve the growth rate. This leaves open a question for efficient computation of weights and constants in these formula. One of the possible approaches for estimating weights and constants is to treat them as parameters and estimate them together with means and variances. This approach will be investigated in future experiments.

Acknowledgment The author would like to thank Leonid Rashevsky, Vaibhava Goel and Peder Olsen for useful discussions and help in preparation of this paper.

7. REFERENCES

[1] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative Training of Subspace Constrained GMMs for Speech Recognition," to be submitted to IEEE Transactions on Speech and Audio Processing.

[2] L.E.Baum and J.A. Eagon, "An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp.360-363, 1967.

[3] A. Gunawardana and W. Byrne, "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," ICASSP, 2002.

[4] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo and A. Nadas, "An inequality for rational functions with applications to some statistical estimation problems", IEEE Trans. Information Theory, Vol. 37, No.1 January 1991

[5] D. Kanevsky, "Growth Transformations for General Functions", RC22919 (W0309-163), September 25, 2003.

[6] Y. Normandin, "An improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition", Proc. ICASSP'91, pp. 537-540, 1991.

[7] R. Schluter, W. Macherey, B. Muler and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition", Speech Communication, Vol. 34, pp.287-310, 2001.

[8] V. Valtchev , P.C. Woodland and S. J. Young, "Lattice-based Discriminative Training for Large Vocabulary Speech Recognition Systems", Speech Communication, Vol. 22, pp. 303-314, 1996.