RC22919 (W0309-163) September 25, 2003
Mathematics

# IBM Research Report

## Growth Transformations for General Functions

**Dimitri Kanevsky**
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

# Growth transformations for general functions

Dimitri Kanevsky

*Plan*



## 1 Introduction

Last decade the new discrimination technique for estimating of parameters became popular. It is based on the transformation formula for continuous parameters [9]. This formula was obtained as approximation of the Baum-Eagon like growth transformation formula for rational functions of discrete parameters that was introduced in [5]. The paper deals mostly with theoretical aspects related to [5]. One of the goal of this paper is to give several proofs for growth transformations for these transformation formula in the case of continuous parameters. The first proof is based on the modification of the basic principle of adding specific constants that was introduced in [5] and that allowed to extend to rational functions Baum-Eagon like growth transformations for polynomial functions. The other proof is based on the lineriazation of the problem for nonlinear functions and computing explicitly the growth estimate for linear forms of Gaussians using a sufficiently large specific constant. In the paper we also give a new proof of the growth of Baum-Eagon like transformation formula for arbitrary objective functions of discrete parameters generalizing [5]. And finally, we derive new transformation formula for continuos parameters case and run simulation experiments to compare growth for different transformation formula.

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

## 2 Growth transformations for rational functions with discrete parameters

Let $R(z) = P_1(z)$, or $R(z) = P_1(z)/P_2(z)$ where $P_1, P_2$ are homogenous polynomials of the same degree $m$ with positive coefficients and $z \in D = \{z_{ij} \geq 0, \sum_j z_{ij} = \sum_{j=1}^{j=m_i} z_{ij} = 1\}$
The following *growth transformation* $z_{ij} \mapsto \hat{z}_{ij}$ was obtained in [5] for sufficiently large $C$.

$$\hat{z}_{ij} = \frac{z_{ij}(\frac{\delta}{\delta z_{ij}} R(z) + C)}{\sum_i z_{ij}(\frac{\delta}{\delta z_{ij}} R(z) + C)} \tag{1}$$

In other words, for sufficiently large $C = C(z)$ the following property holds: $R(\hat{z}) > R(z)$ if $\hat{z} \neq z$

## 3 Linearization

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

Let $F : z \in R^n \to R^1$ be some function. We tell that $F$ is *good* at $(z, i)$ if there exists such a small ball $V = V_z(\epsilon) = \{z\prime | |z\prime - z| < \epsilon\}$ at a center $z$ that for any $z\prime \in V$ the following holds: $F(z\prime) - F(z) = \sum_i \frac{\delta F(z)}{\delta z_i}(z_i\prime - z_i) + O(|z\prime - z|^{1+\delta})$, where $\delta > 0$ and $\frac{\delta F(z)}{\delta z_i} \neq 0$. For example, $F$ is *good* at $(z, i)$ if it has all derivatives of a second order at $z$ and its derivative of the first order by $z_i$ is not equal to zero at $z$.We also will tell that $F$ is good at $z$ if it is good at $(z, i)$ for some $i$.

**Lemma 1** *Let*

$$F(z) = F(\{u_j\}) = F(\{g_j(z)\}, j = 1, ..m \tag{2}$$

*be a function that can be represented as a composite of a system of $m$ functions $u_j = g_j(z)$ where $z$ varies in some real vector space $R^n$ of dimension $n$. Let, futher, $L(z\prime) = L(\{g_i(z\prime)\}) = \sum_j \frac{\delta F(\{u_j\})}{\delta u_j} g_j(z\prime)$ where $\frac{\delta F(\{u_j\})}{\delta u_j}$ is taken at $u_j = g_j(z)$ and $z\prime \in R^n$ . Let $F$ and $L$ be good at $(z, i)$. Let $T_\epsilon$ be a family of transformations $R^n \to R^n$ that factors through the transformation $dF : z \in R^n \to (\{\frac{\delta F(z)}{\delta z_j}\}) \in R^m$, i.e. there exists a family of map $G_\epsilon : R^m \to R^n$, such that $T_\epsilon = G_\epsilon dF$ . Assume also that $T_\epsilon(z) \to z$ if $\epsilon \to 0$ and $T_\epsilon(z)_i \neq z_i$ for some $i$ .Then there exists such a small $V_z(\epsilon)$ that $T_\epsilon$ is growth for sufficiently small $\epsilon$ for $F$ at $z$ iff $T_\epsilon$ is growth for $L$ at $z$.*

.

*Proof*
First, from the definition of $L$ we have $\frac{\delta F(z)}{\delta z_k} = \sum_j \frac{\delta F(\{u_j\})}{\delta u_j} \frac{\delta g_j(z)}{\delta z_k} = \frac{\delta L(z)}{\delta z_k}$

Next we have: $F(z\prime) - F(z) = \sum_i \frac{\delta F(z)}{\delta z_i}(z_i\prime - z_i) + O(\alpha^{1+\delta}) = \sum_i \frac{\delta L(z)}{\delta z_i}(z_i\prime - z_i) + O(\alpha^{1+\delta_1}) = L(z\prime) - L(z) + O(\alpha^{1+\delta_2})$, where $\alpha = |z\prime - z|$, $\delta_1 > 0$ and $\delta_2 > 0$, $z\prime = T_\epsilon(z)$ and $\epsilon$ is sufficiently small. Therefore for sufficientlay small $\epsilon$ $F(z\prime) - F(z) > 0$ iff $L(z\prime) - L(z) > 0$.

# 4 Proof of transformation formula for general functions

. Here we give a different proof of (1). This proof generalize the statement to any function that allows linearization, i.e. that have derivatives of the second order. Therefore we assume now that $R(z)$ is arbitrary function that has all derivatives of the second order.

According to the linearization principle, we can assume that $R(z) = l(z) = \sum a_{ij}z_{ij}$ is a linear form.

Than the transformation formula for $l(x)$ is the following:

$$\hat{z}_{ij} = \frac{a_{ij}z_{ij} + Cz_{ij}}{l(z) + C} \tag{3}$$

We need to show that

$$l(\hat{z}) \geq l(z) \tag{4}$$

It is suffcient to proove 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 sub-index $i$ in the fortcoming 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:

$$l(\hat{z}_j) = \frac{l_2(z) + Cl(z)}{l(z) + C} \tag{5}$$

Where

$$l_2(z) := \sum_j a_j^2 z_j \tag{6}$$

We need to prove that the following

**Lemma 2**

$$l_2(z) \geq l(z)^2 \tag{7}$$

*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 proove:

$$\sum_{j=1}^{j=n-1} [a_j^2 x_j + a_n^2(1 - z\prime)] \geq \sum_{j=1}^{j=n-1} (a_j - a_n)^2 z_j^2 + 2\sum_{j=1}^{j=n-1} (a_j - a_n)a_n z\prime^2 + a_n^2 \tag{8}$$

We will proove the above formula by prooving for every fixed $j$

$$(a_j^2 - a_n^2)z_j \geq (a_j - a_n)^2 z_j^2 + 2(a_j - a_n)a_n z_j \tag{9}$$

If $(a_j - a_n)z_j \neq 0$ then the above inequality is equivalent to

$$a_j + a_n > (a_j + a_n)z_j \tag{10}$$

The above ineqaultiy 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 (7) we have the following inequalities. $l_2(z) + Cl(z) \geq l(z)^2 + Cl(z)$ $l(\hat{z}) = \frac{l_2(z)+Cl(z)}{l(z)+C} \geq \frac{l(z)^2+Cl(z)}{l(z)+C}$ if $l(z) + C > 0$ and $l(\hat{z}) = \frac{l_2(z)+Cl(z)}{l(z)+C} \leq \frac{l(z)^2+Cl(z)}{l(z)+C}$ if $l(z) + C < 0$ This proves the statement.

The following theorem is a generalization of the statement that was given in [7] for growth transformations for analytic functions.

**Theorem 1** *Let $F(z)$ is a function that is defined over $D = \{z_{ij} \geq 0, \sum z_{ij} = 1\}$. Let $F$ be good at $z \in D$. Let*

$$\hat{z}_{ij} = \frac{z_{ij}(\frac{\delta}{\delta z_{ij}} F(z) + C)}{\sum_i z_{ij}(\frac{\delta F(z)}{\delta z_{ij}} F(z) + C)} \tag{11}$$

*And let $\hat{z} \neq z$ for sufficiently large $|C|$. Then $F(\hat{z}) > F(z)$ for sufficiently large positive $C$ and $F(\hat{z}) < F(z)$ for sufficiently small negative $C$.*

*Proof* It follows immediately from linerization principle and the previous lemma.

# 5 New optimization principle

Let us consider the following polynomial: $P(X) = \sum c_\nu X_\nu^{\nu_i}$ where $c_\nu$ are coefficients in a polynomial, $X_\nu^{n_\nu} = \prod X_i^{n_i}$, $X_i, i = 1, ...l$ are variables , $\nu = \{i_1, i_2, ...\}$ is a multi-index and $n_\nu = \{n_{i_1}, n_{i_2}, ..., n_{i_l}\}$ . Let $x = (x_1, x_2, ..., x_l)$ , $\tilde{x} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_l)$ be some points with all $x_i, \tilde{x}_i \geq 0$ and such that $\sum x_i = \sum \tilde{x}_i > 0$. Let $L_P(x, \tilde{x}) = \sum x_i c_i \log \tilde{x}_i$, where $c_i = \frac{\delta}{\delta x_i} P(\{x_i\})$. We call $L_P$ associated with $P$ at $x$ and $\tilde{x}$. It is well known (see [3]) that if all coefficients in $P$ are non negative then the inequality $L_P(x, \tilde{x}) > L_P(x, x)$ implies the inequality $P(\tilde{x}) > P(x)$ .
If the polynomial $P$ does not have all coefficients positive one can use the following statement

**Proposition 1** *Let $D(x)$ be a polynomial and $\tilde{x}$ another point such that the following holds:*

$$K(x) = P(x) + D(x) \tag{12}$$

*has all coefficients non-negative,*

$$D(\tilde{x}) \leq D(x) \tag{13}$$

*and*

$$L_K(x, \tilde{x}) > L_K(x, x) \tag{14}$$

*Then $x \mapsto \tilde{x}$ is a growing transformation, i.e. $P(\tilde{x}) > P(x)$.*

*Proof*
$K(\tilde{x}) > K(x)$ because of (12) and (14). The final statement now follows from (13).

This generalizes the principle in [5] in which it was assumed that $D(x)$ is a constant for all $x$ in a probability domain. If $x$ depends on some continuous parameters (12) and (13) conditions give rise to some equations for these parameters. Solving these equations can lead to new optimization procedures. Later we use these considerations to deduct an iterative optimization formula for continuous parameters.

# 6 Heuristic iterative formula for continuous parameters

In this section we derive a heuristic re-estimation formula for models with continuous parameters, using (1). These formula were initially obtained in [6], [9].
Let $Y = \{y_1, ...y_K\}$ denotes a training data, where $y_i$ are real numbers. Let $x_{ij} = N(y_i, \mu_j, \sigma_j), i = 1, ...K$ be one dimensional Gaussian densities.
Let

$$R(\{\mu_j, \sigma_j\}) = R(\{x_{ij}\}) = \frac{P_1(\{x_{ij}\})}{P_2(\{x_{ij}\})} \tag{15}$$

be a rational function where either $P_1, P_2$ are homogenous polynomials of $x_{ij}$ of the same degree $m$ with positive coefficients or $P_2 = 1$.

We want to find re-estimation formula that resolves the following

## 6.1 Problem

Find

$$Arg \max_{\{\mu_j, \sigma_j\}} R(\{\mu_j, \sigma_j\}) \tag{16}$$

We need to introduce some notations to derive the heuristic formula for a growth transformation for the problem (16). We will follow [8] approach in derivation of heuristic formula.

Let partition a real axis (domain of Gaussian density) into three non-overlapping intervals: $I_1 = (-\infty, \mu_j - \nu)$
$I_2 = [\mu_j - \nu, \mu_j + \nu]$
$I_3 = (\mu_j + \nu, +\infty)$
Let partition $I_2$ in $T$ non-overlapping non-zero sub-intervals $\Delta_k$ of length $h_k \leq h$. Choose $\nu$ so large that all points of the training data $y_1, ...y_K \in Y$ fall in the second segment $I_2$. Let $x_1 \in \Delta_1, x_2 \in \Delta_2, ...x_T \in \Delta_T$ be some points in sub-intervals $\Delta_i$. Let $\Delta_k$ are chosen so small that each $\Delta_k$ contains not more than one $y_i$ from the sampling data $Y$. Let us denote a set of all $\Delta_k$ each of which contain some $y_i$ from the sampling data $Y$ as $\tilde{\Delta}$. Let us change $x_i$ and enumerate $x_i$, $y_j$, $\Delta_k$ in such a way that $x_i = y_i \in \Delta_i$ if $y_i$ belonged some $\Delta_j \in \tilde{\Delta}$.
Let $I = \{k : \Delta_k \in \tilde{\Delta}\}$ denote a set of all indexes for $\Delta_k \in \tilde{\Delta}$. Let denote by $W = W(I, Y, X, \{\Delta_i\})$ a system that contains the set of indexes $I$, the training data $Y$, the set of points $X = \{x_1, x_2, ...x_T\}$, and the set sub-intervals $\{\Delta_i\}$).
It is clear that if $\nu$ grows than $T$ also grows but the size of $I$ depends only on the size of training data $Y$ that does not changes with growth of $T$. Let now all $h_k = h$ and let us define

$$a_{ij} = \frac{N(x_i, \mu_j, \sigma_j)h}{\sum_i N(x_i, \mu_j, \sigma_j)h} \tag{17}$$

For any $y$ and sub-interval $\Delta$ containing $x_i$ the following holds:

$$lim_{h \to 0} x_i = y$$
$$lim_{h \to 0} N(x_i, \mu_j, \sigma_j) = N(y, \mu_j, \sigma_j)$$
$$lim_{h \to 0, \nu \to \infty} \sum_i N(x_i, \mu_j, \sigma_j)h = 1$$

Let us consider the following "discrete approximation" procedure. We substitute $N(x_i, \mu_j, \sigma_j)$ with $\{a_{ij}\}$ in (16).

$$R(\{\mu_j, \sigma_j\}) = R(\{x_{ij}\}) \to R(\{a_{ij}\})$$

Then

$$lim_{h\to 0,\nu\to\infty} R(\{a_{ij}\}) = R(\{x_{ij}\})$$

Consider the following discrete optimization problem:

$$Arg \max_{\{a_{ij}\}} R(\{a_{ij}\}) \tag{18}$$

Let consider the following growth transformation for the problem (18) $\{a_{ij}\} \mapsto \{\hat{a}_{ij}\}$ .

$$\hat{a}_{ij} = \frac{a_{ij}(\frac{\delta}{\delta a_{ij}} R(\{a_{ij}\}) + C)}{\sum_i a_{ij}(\frac{\delta}{\delta a_{ij}} R(\{a_{ij}\}) + C)} \tag{19}$$

Let obtain new continuous parameters via the following approximation:

$$\hat{\mu_j} = lim_{h\to 0,\nu\to\infty} \sum_i \hat{a}_{ij} x_i \tag{20}$$

$$\hat{\sigma_j}^2 = lim_{h\to 0,\nu\to\infty} \sum_i \hat{a}_{ij} (x_i - \hat{\mu_j})^2 \tag{21}$$

We can compute $\hat{\mu}_j$ and $\hat{\sigma}_j^2$ using the following equalities.

$$lim_{h\to 0,\nu\to\infty} \sum_i a_{ij} x_i = \mu_j \tag{22}$$

$$lim_{h\to 0,\nu\to\infty} \sum_i a_{ij} x_i^2 = \mu_j^2 + {\sigma_j}^2 \tag{23}$$

Let $c_{ij} = \frac{\delta}{\delta x_{ij}} R(\{x_{ij}\}$. Using (22) and (23) we get the following transformation formula:

$$\hat{\mu}_j = \hat{\mu}_j(C) = \frac{\sum_{i\in I} x_{ij} c_{ij} x_i + C\mu_j}{\sum_{i\in I} x_{ij} c_{ij} + C} \tag{24}$$

$$\hat{\sigma}_j^2 = \hat{\sigma}_j(C)^2 = \frac{\sum_{i\in I} x_{ij} c_{ij} x_i^2 + C(\mu_j^2 + {\sigma_j}^2)}{\sum_{i\in I} x_{ij} c_{ij} + C} - \hat{\mu}_j^2 \tag{25}$$

The problem with this heuristic development is that a constant $C = C(a_{ij})$ is obtained from a discrete formulae (19) and depends on $a_{ij}$, i.e. depends on $h$. When $h \to 0$ then $C \to \infty$ in (24) and (25). This is shown in the Appendix.
In practice iterative algorithms that are based on these formula provided good incremental growth for discrimination objective functions (that involve functions like (2)). The goal of the next chapter is to prove the following statement.

**Theorem 2** *For sufficiently large C the map (24),(25)* $\{\mu_j, \sigma_j\} \mapsto \{\hat{\mu}_j, \hat{\sigma}_j\}$ *is growth transformation, i.e.* $R(\{\hat{x}_{ij}\}) > R(\{x_{ij}\})$ *if* $\{\hat{x}_{ij}\} \neq \{x_{ij}\}$.

*Remark:* Vaibhava Goel informed me that Axelrod [1], [2], has recently proposed another proof of existance of C that ensures validity of the MMIE auxiliary function as formulated by Gunawardana et.al. [4]. His derivation applies in general to density functions that obey certain smoothness constraints around the current parameter value.

# 7 Proof of Theorem 2

We first prove the variant of this theorem for polynomials. Then we deduct the statement for rational functions. Let $f(x, \mu, \sigma)$ be density (e.g. $N(x, \mu, \sigma)$). Let $P(x_{ij})$ be a homogenous polynomial in $x_{ij}$ of $\deg m$ where $i \in I$. Let us consider a function $P(f(x_i, , \mu_j, \sigma_j))$ that is obtained from $P(x_{ij})$ by substituting $x_{ij}$ with $f(x_i, \mu_j, \sigma_j)$. Here $x_i$ are of values from a sample of training data $i \in I$ and $\mu_j$, $\sigma_j$ are parameters . Consider the following problem

$$Arg \max_{\{\mu_j, \sigma_j\}} P(f(x_i, \mu_j, \sigma_j)) \tag{26}$$

If the polynomial (26) has all coefficients positive then one can generate growth transform as described in [3] . Otherwise, we need to reduces the original problem to the new one that involves only the polynomial with the positive coefficients. Our proof consists of the following

## 7.1 Steps:

1. First we consider a discrete variant of the continuous problem that associate with a transformation some large constant $C$.
2. We will use the new principle and constraints (12), (13). introduce a system of equations for $\mu, \sigma$. The coefficients in these equations will be "guessed" from formula (24),(25) that were obtained via heuristics limit procedure. Therefore solutions of this system of equations will be exactly formulae (24),(25).

## 7.2 Discreditation

Here we consider new notation (assuming also notation of Section 4).
*Notation*

Let $z_k, k \in \{1, ...T\}$ be unknowns whose values belong to a domain that will be described later.

Let set: $A_{ij} = A_{ij}(z_i) = f(x_i, \mu_j, \sigma_j)z_i$. and
$\hat{A}_{ij} = \hat{A}_{ij}(z_i) = f(x_i, \hat{\mu}_j, \hat{\sigma}_j)z_i$.

We also set
$f_{ij} = f(x_i, \mu_j, \sigma_j) = N(x_i, \mu_j, \sigma_j)$
$\hat{f}_{ij} = f(x_i, \hat{\mu}_j, \hat{\sigma}_j) = N(x_i, \hat{\mu}_j, \hat{\sigma}_j)$

For any $\mu, \sigma, \hat{\mu}, \hat{\sigma}$ let define the following set $E = E_C(I, Y, X, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$ of equations and constrains for $z_k$.

$$z_k = z_{k'} \tag{27}$$

if $k, k' \in I$. We denote $z_k = z$ if $k \in I$

$$z_k >= 0 \tag{28}$$

$$\sum_{i=1} A_{ij} = 1 \tag{29}$$

$$\sum_{i=1} \hat{A}_{ij} = 1 \tag{30}$$

$$\sum_{i=1} A_{ij} x_i = \mu_j \tag{31}$$

$$\sum_{i=1} A_{ij} x^2{}_i = \mu_j^2 + \sigma_j^2 \tag{32}$$

It is clear that in the system $E_C = E_C(I, X, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$ the data $I, X$ can be constructed from $W = W(I, Y, X, \{\Delta_i\}))$. We therefore sometime will denote $E_C(I, Y, X, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$ as $E_C = E_C(W, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$. The value of the introduced system of the equations and inequalities can be seen from the following statement:

**Lemma 4** *Let $\hat{\mu}_j = \hat{\mu}_j(C)$ and $\hat{\sigma}_j = \hat{\sigma}_j(C)$ are defined as in (24), (25) for some $I, Y, \mu, \sigma$. Then there exist such large $C$ and such $X$ containing $Y$ that if the system $E_C$ has a solution, then $\{\mu_j, \sigma_j\} \mapsto \{\hat{\mu}_j, \hat{\sigma}_j\}$ is growth transformation*

*Proof*

Let fix any positive value of $z$. First, we have the following implication: if $P(\{\hat{f}_{i,j}\}) > P(\{f_{i,j}\})$ then

$$P(\{\hat{A}_{i,j}\}) = P(\{\hat{f}_{i,j}\})z^m > P(\{A_{i,j}\}) = P(\{f_{i,j}\})z^m \tag{33}$$

This follows from the fact that $P()$ is a homogenous polynomial and that only those $A_{i,j}$ are considered in $P()$ for which $\hat{A}_{i,j} = \hat{f}_{i,j}z$ and $A_{i,j} = f_{i,j}z$. (In other words, $I$ is chosen in such a way that for any $x_{ij}$ in $P$ $i \in I$). If $\{A_{i,j}\} \mapsto \{\hat{A}_{i,j}\}$ is a growth transformation for $P(\{A_{i,j}\})$, i.e. the inequality (33) holds, then it is equivalent to the following inequality:

$$P(\{\hat{A}_{i,j}\}) + C(\sum_{i=1,T}\{\hat{A}_{i,j}\})^m > P(\{A_{i,j}\}) + C(\sum_{i=1,T}\{A_{i,j}\})^m \tag{34}$$

for any $C$ since $(\sum_i\{\hat{A}_{i,j}\}) = \sum_i\{A_{i,j}\} = 1$ is constant (by (29) and (30)). Choosing $C$ sufficiently large we get all coefficients in $P_C = P(\{A_{i,j}\}) + C(\sum_i\{A_{i,j}\})^m$ positive. Let $Z = Z(C) = \{z_k\}$ be some solution of $E_C(W, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$. Then (34) is the consequence of the following two facts:

**Fact 1** Let $C_{ij}$ are computed as follows:

$$C_{ij} = \frac{\delta}{\delta A_{ij}} P(A_{ij}) \tag{35}$$

Then (33) holds if the following inequality holds:

$$\sum_{i=1,T}(A_{ij}C_{ij} + CA_{ij})\log\hat{A}_{ij} > \sum_{i=1,T}(A_{ij}C_{ij} + CA_{ij})\log A_{ij} \tag{36}$$

This is a consequence of Jensen inequality for concave functions [3].

**Fact 2** Let $g(\mu_j, \sigma_j) = \sum_{i=1,T}(A_{ij}C_{ij} + CA_{ij})\log f_{ij}$ where $f_{ij} = N(x_i, \mu_j, \sigma_j)$. and let

$$(\hat{\mu}_j, \hat{\sigma}_j) = Arg \max_{\{\mu_j, \sigma_j\}} g(\mu_j, \sigma_j) \tag{37}$$

Then $(\hat{\mu}_j, \hat{\sigma}_j)$ equals (24), (25) with some replacement for $C$.

We start from solving the problem (37).

This problem (37) is the problem of maximization of concave function and can be resolved easily via the following standard methods.

$$\frac{\delta}{\delta s_j} g(\mu_j, \sigma_j)|_{\hat{\mu}_j, \hat{\sigma}_j} = 0 \tag{38}$$

where $s_j = \mu_j$ or $\sigma_j$. The (38) leads to the following equations for $\hat{\mu}_j$ and $\hat{\sigma}_j$.

$$\sum_i (A_{ij}C_{ij} + CA_{ij})(x_i - \hat{\mu}_j) = 0 \tag{39}$$

$$\sum_i (A_{ij}C_{ij} + CA_{ij})(-1 + \frac{(x_i - \hat{\mu}_j)^2}{\hat{\sigma}_j^2}) = 0 \tag{40}$$

The solution of (39) is

$$\hat{\mu}_j = \frac{\sum_i A_{ij}C_{ij}x_i + C\sum_i A_{ij}x_i}{\sum_i A_{ij}C_{ij} + C\sum_i A_i} \tag{41}$$

And the solution of (40) is

$$\hat{\sigma_j}^2 = \frac{\sum_i A_{ij}C_{ij}x_i^2 + C\sum_i A_{ij}x_i^2}{\sum_i A_{ij}C_{ij} + C\sum_i A_i} - \hat{\mu}_j^2 \tag{42}$$

Since $C_{ij} = 0$ for $i$ not in $I$ $\sum_i A_{ij}C_{ij}x_i = \sum_{i\in I} A_{ij}C_{ij}x_i$. This allows to re-write (41) and as follows.

$$\hat{\mu}_j = \frac{\sum_{i\in I} A_{ij}C_{ij}x_i + C\sum_{i=1,T} A_{ij}x_i}{\sum_{i\in I} A_{ij}C_{ij} + C\sum_{i=1,T} A_{ij}} \tag{43}$$

Similarly one can re-write (42) as follows

$$\hat{\sigma_j}^2 = \frac{\sum_{i\in I} A_{ij}C_{ij}x_i^2 + C\sum_{i=1,T} A_{ij}x_i^2}{\sum_{i\in I} A_{ij}C_{ij} + C\sum_{i=1,T} A_{ij}} - \hat{\mu}_j^2 \tag{44}$$

Using (29), (30), (31) and (32) we get the following values:

$$\hat{\mu}_j = \frac{\sum_{i\in I} A_{ij}C_{ij}x_i + C\mu_j}{\sum_{i\in I} A_{ij}C_{ij} + C} \tag{45}$$

And

$$\hat{\sigma_j}^2 = \frac{\sum_{i\in I} A_{ij}C_{ij}x_i^2 + C(\mu_j^2 + \sigma_j^2)}{\sum_{i\in I} A_{ij}C_{ij} + C} - \hat{\mu}_j^2 \tag{46}$$

Replacing $A_{ij}$ with $f_{ij}$ and $C$ with $C/z$ gives (24), (25). The theorem 2 will follow from the following

**Lemma 5** *For any $C, \mu, \sigma, \hat{\mu}, \hat{\sigma}, Y, I$ there exists $X$ containing $Y$ such that $E_C$ has non-empty solution $Z = \{z_k\}$.*

**Proof**
The proof consists of the several steps.
**Step 1**
Given conditions of Lemma 2 and arbitrary positive constant $d$ one can choose $W$ such that length of all $\Delta_k = h$ and $h$ is so small and $T$ is so large that the following system of equations and inequalities hold for $z_k = h$:

$$\sum_{i=1} A_{ij} = 1 + d_1(h^{1+\delta})$$

$$\sum_{i=1} \hat{A}_{ij} = 1 + d_2(h^{1+\delta})$$

$$\sum_{i=1} A_{ij}x_i = \mu_j + d_3(h^{1+\delta})$$

$$\sum_{i=1} A_{ij}x_i^2 = \mu_j^2 + \sigma_j^2 + d_4(h^{1+\delta})$$

$$h + d(h^{1+\delta}) > 0 \tag{47}$$

where $|d_i| < d$ and $\delta > 0$.

**Step 2**

One can choose some 4 different $\hat{i} = \{1 < i_1, i_2, i_3, i_4 <= T\}$ outside $I$ and $x_{i_r}, r = 1, 2, 3, 4$ such that linear independent columns in the system of equations are generated in Step 1

$$\begin{matrix} f_{i_1 j} & f_{i_2 j} & f_{i_3 j} & f_{i_4 j} \\ \hat{f}_{i_1 j} & \hat{f}_{i_2 j} & \hat{f}_{i_3 j} & \hat{f}_{i_4 j} \\ f_{i_1 j} x_{i_1} & f_{i_2 j} x_{i_2} & f_{i_3 j} x_{i_3} & f_{i_4 j} x_{i_3} \\ f_{i_1 j} x_{i_1}^2 & f_{i_2 j} x_{i_2}^2 & f_{i_3 j} x_{i_3}^2 & f_{i_4 j} x_{i_3}^2 \end{matrix}$$

This follows from the fact that determinant of this matrix is a polynomial of $x_j$ and it defines a variety of co-dimension one on some space. Slightly varying $x_i$ outside of this variety one can obtain $x_i$ that belong intervals that are defined by $\Delta_k$ and that determinant of this matrix is non-zero.

**Step 3** Let replace unknown $z_i$ with $i \in \hat{i}$ by $h + \epsilon_i$. Then the system of equalities in (47) gives rise to the following system of equations.

$$\sum_{r=1,4} f_{i_r j} \epsilon_{i_r} = -d_1(h^{1+\delta})$$

$$\sum_{r=1,4} \hat{f}_{i_r} \epsilon_{i_r} = -d_2(h^{1+\delta})$$

$$\sum_{r=1,4} f_{i_r j} x_i \epsilon_{i_r} = -d_3(h^{1+\delta})$$

$$\sum_{r=1,4} f_{i_r j} x_i^2 \epsilon_{i_r} = -d_4(h^{1+\delta})$$

This system of equation is solvable since the determinant of the system is non-zero. The solutions of this system satisfy inequalities $|\epsilon_{i_r}| < d_5(h^{1+\delta})$ for some large $d_5$ as can be seen from explicit solutions of this system. If $d_5 > d$, let put $d = d_5$ and choose $h$ sufficiently small and $T$ so large that (47) holds. Then one can choose the following solution of the system of equations $E_C$: $z_k = h$ if $k$ does not belong to $\hat{i}$ and $z_{i_r} = h + \epsilon_{i_r} > 0$ .

Q.E.D.

The theorem 2 for rational functions now follows by standard reduction of the rational function $R$ to a polynomial $P_1 - kP_2$ for some coefficient $k$ (see Appendix 1) and the fact that coefficients (35) for polynomials proportional to coefficients $c_{ij}$ for rational functions in (24), (25).

### 7.3 Generalization

In the notation of 5.2 for any $\mu, \sigma, \hat{\mu}, \hat{\sigma}$ let us define the following general set $D_C = D_C(I, Y, X, \mu, \sigma, \hat{\mu}, \hat{\sigma}, \{z_k\})$ of equations and constrains for $z_k$.

$$z_k = z_{k'} \tag{48}$$

if $k, k' \in I$. We denote $z_k = z$ if $k \in I$

$$z_k >= 0 \tag{49}$$

$$\sum_{i=1} A_{ij} = \sum_{i=1} \hat{A}_{ij} \tag{50}$$

**Lemma 6** *Let $\hat{\mu}_j = \hat{\mu}_j(C)$ and $\hat{\sigma}_j = \hat{\sigma}_j(C)$ are defined as in (43), (44) for some $I, Y, \mu, \sigma$. Then there exist such large $C$ and such $X$ containing $Y$ that if the system $D_C$ has a solution, then $\{\mu_j, \sigma_j\} \mapsto \{\hat{\mu}_j, \hat{\sigma}_j\}$ that corresponds this solution is growth transformation.*

*Proof*
In the proof of this lemma we can follow the proof of the lemma 1 until (44).

# 8 Another proof of growth transformations for general functions with continuous parameters

Let $R$ in (15) be a real function. For simplicity of the notation we consider the transformation (24), (25), only for a single pair of variables $\mu, \sigma$, i.e. $R(\mu, \sigma) = R(N_i)$, where

$$N_i = \frac{1}{(2\pi)^{1/2}\sigma} e^{-(y_i-\mu)^2/2\sigma^2} \tag{51}$$

We also use the notation $c_i = N_i \frac{\delta R}{\delta N_i}$ and

$$\hat{N}_i = \frac{1}{(2\pi)^{1/2}\hat{\sigma}} e^{-(y_i-\hat{\mu})^2/2\hat{\sigma}^2} \tag{52}$$

Let write transformation formula (24), (25) as

$$\hat{\mu} = \hat{\mu}(C) = \frac{\sum_{i\in I} c_i y_i + C\mu}{\sum_{i\in I} c_i + C} \tag{53}$$

$$\hat{\sigma}^2 = \hat{\sigma}(C)^2 = \frac{\sum_{i\in I} c_i y_i^2 + C(\mu^2+\sigma^2)}{\sum_{i\in I} c_i + C} - \hat{\mu}^2 \tag{54}$$

Now we can formulate a theorem that extends applicability of transformation formula (24), (25) to general functions.

**Theorem 3** *Let $\mu, \sigma$ be such that*

$$\sum c_j(y_j - \mu) \neq 0 \tag{55}$$

*or*

$$\sum c_j[(y_j-\mu)^2 - \sigma^2] \neq 0 \tag{56}$$

*Let $R(\mu, \sigma) = R(\{N_i\})$, $i = 1...m$, be good at $\mu, \sigma$. Then for sufficiently large $C$*

$$R(\{\hat{N}_i\}) - R(\{N_i\}) = T/C + o(1/C^2) \tag{57}$$

*Where*

$$T = \frac{1}{\sigma^2}\{\frac{\{\sum c_j[(y_j-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + [\sum c_j(y_j-\mu)]^2\} > 0 \tag{58}$$

*In other words, $R(\{\hat{N}_i\})$ grows proportionaly to $1/C$ for sufficiently large $C$.*

*Proof*
We will prove the theorem via linearization. According to the linearization principle, we can assume that $R((\mu, \sigma) = l(\mu, \sigma) := l(\{N_i\}) := \sum_{i=1}^{i=m} a_i N_i$. Let denote also $l(\hat{\mu}, \hat{\sigma}) := l(\{\hat{N}_i\}) := \sum_{i=1}^{i=m} a_i \hat{N}_i$

We consider the following transformation formula

$$\hat{\mu} = \hat{\mu}(C) = \frac{\sum_{j=1}^{j=m} c_j y_j + C\mu}{\sum_{j=1}^{j=m} c_j + C} \quad (59)$$

where $c_j = a_j N_j$.

$$\hat{\sigma}^2 = \hat{\sigma}(C)^2 = \frac{\sum_{j=1}^{j=m} c_j y_j^2 + C(\mu^2+\sigma^2)}{\sum_{j=1}^{j=m} c_j + C} - \hat{\mu}^2 \quad (60)$$

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 precision $1/C^2$.

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

$$\sim (\frac{1}{C}\sum_{j=1}^{j=m} c_j y_j + \mu)(1 - \frac{\sum c_j}{C}) \sim \mu + \frac{1}{C}(\sum_{j=1}^{j=m} c_j y_j - \mu \sum_{j=1}^{j=m} c_j) \quad (61)$$

$$\hat{\mu} \sim \mu + \frac{\sum_{j=1}^{j=m}[c_j(y_j-\mu)]}{C} \quad (62)$$

Let compute $\hat{\sigma}^2$ using (60)s

$$\frac{\sum_{j=1}^{j=m} c_j y_j^2 + C(\mu^2+\sigma^2)}{\sum_{j=1}^{j=m} c_j + C} \sim$$

$$\sim (\frac{\sum_{j=1}^{j=m} c_j y_j^2}{C} + \mu^2 + \sigma^2)(1 - \frac{\sum_{j=1}^{j=m} c_j}{C}) \sim$$

$$\sim \mu^2 + \sigma^2 + \frac{1}{C}[\sum_{j=1}^{j=m} c_j y_j^2 - (\mu^2+\sigma^2)\sum_{j=1}^{j=m} c_j] \quad (63)$$

$$\hat{\mu}^2 \sim \mu^2 + \frac{2\mu}{C}\sum_{j=1}^{j=m} c_j(y_j-\mu) \quad (64)$$

This gives

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

$$= \sigma^2 + \frac{1}{C}[\sum_{j=1}^{j=m} c_j y_j^2 - (\mu^2+\sigma^2)\sum_{j=1}^{j=m} c_j - 2\mu(\sum_{j=1}^{j=m} c_j(y_j-\mu)] \quad (65)$$

And finally

$$\hat{\sigma}^2 \sim \sigma^2 + \frac{\sum[(y_j-\mu)^2 - \sigma^2]c_j}{C} \quad (66)$$

$$(y_i-\hat{\mu})^2/\hat{\sigma}^2 \sim \frac{1}{\sigma^2}[(y_i-\mu)^2 - \frac{2(y_i-\mu)\sum_{j=1}^{j=m} c_j(y_j-\mu)}{C}]\{1 - \frac{\sum_{j=1}^{j=m} c_j[(y_j-\mu)^2-\sigma^2]}{\sigma^2 C}\} \sim$$

$$\sim \frac{(y_i-\mu)^2}{\sigma^2} - \frac{1}{C\sigma^2}\{\frac{(y_i-\mu)^2}{\sigma^2}\sum[(y_j-\mu)^2+\sigma^2]c_j + 2(y_i-\mu)\sum(y_j-\mu)c_j\} \quad (67)$$

$$\hat{N}_i \sim \frac{1}{(2\pi)^{1/2}\hat{\sigma}} e^{\frac{-(y_i-\mu)^2}{2\sigma^2}+\frac{A}{C\sigma^2}} \tag{68}$$

Where

$$A_i = \frac{(y_i-\mu)^2}{2\sigma^2}\sum[(y_j-\mu)^2-\sigma^2]c_j + (y_i-\mu)\sum(y_j-\mu)c_j \tag{69}$$

Continue this we have

$$\hat{N}_i \sim K e^{\frac{-(y_i-\mu)^2}{2\sigma^2}}(1+\frac{A_i}{C\sigma^2}) \tag{70}$$

Where

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

$$1/\hat{\sigma} \sim \frac{1}{\sigma}\{1 - \frac{\sum c_i[(y_i-\mu)^2-\sigma^2]}{2\sigma^2 C}\} \tag{71}$$

$$(1+\frac{A_i}{C\sigma^2})\{1-\frac{\sum c_i[(y_i-\mu)^2-\sigma^2]}{2\sigma^2 C}\} \sim$$

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

$$\sim 1+\frac{1}{C\sigma^2}\{[\frac{(y_i-\mu)^2}{2\sigma^2}-1/2]\sum[(y_i-\mu)^2-\sigma^2]c_j+(y_i-\mu)\sum(y_j-\mu)c_j]\} \tag{72}$$

$$\sim 1+\frac{B_i}{C\sigma^2} \tag{73}$$

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

Using the last equalities we get

$$\hat{N}_i = N_i + \frac{B_i}{C\sigma^2}N_i \tag{74}$$

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

$$l(\{\hat{N}_i\}) = l(\{N_i\}) + \frac{l(\{B_i N_i\})}{C\sigma^2} \tag{75}$$

and

$$L(\{B_i N_i\}) = \sum a_i N_i\{[\frac{(y_i-\mu)^2}{2\sigma^2}-1/2]\sum c_j[(y_j-\mu)^2-\sigma^2]+(y_i-\mu)\sum c_j(y_j-\mu)\} \tag{76}$$

$$= \sum c_i\{[\frac{(y_i-\mu)^2}{2\sigma^2}-1/2]\}\sum c_j[(y_j-\mu)^2-\sigma^2]+(y_i-\mu)\sum c_j(y_j-\mu)\} \tag{77}$$

$$= \frac{\{\sum c_j[(y_j-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + [\sum c_j(y_j-\mu)]^2 \tag{78}$$

$$l(\{\hat{N}_i\}) - l(\{N_i\}) \sim \frac{1}{C\sigma^2}\{\frac{\{\sum c_j[(y_j-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + [\sum c_j(y_j-\mu)]^2\} \tag{79}$$

# 9 Another Baum Growth Transformation formulae for general "good" functions with continuos parameters

In this section we derive a new re-estimation formula for models with continuous parameters for general functions that have some good properties that will be specified later.
We refer to this transformation as modified Baum (and refer to the previous transformation as standard Baum).
Let $Y = \{y_{ij}\}$ denotes a training data, where $y_{ij}$ are real numbers. Let $N_{ij} = N(y_{ij}, \mu_j, \sigma_j), i = 1, ...k, j = 1...m$ be one dimensional Gaussian densities. Let $f(\{N_i\}) = f(\{\mu_i, \sigma_i\})$ be a general function from Gaussians $N_{ij}$. We derive the formula under assumptions that all $0 \leq \mu_i \leq D_i, 0 \leq \sigma_i \leq E_i$. Then we can introduce slack variables $\mu\prime \geq 0$, $\sigma\prime \geq 0$ such that $\mu + \mu\prime = 1$, $\sigma + \sigma\prime = 1$. Then we can compute updates for $\mu$ and $\sigma$ using (1). This gives rise to the following growth transformations:

$$\hat{\mu}_j = D_j \mu_j \frac{\sum_{i \in \{1...k\}} \frac{\delta f(\{N_{ij}\})}{\delta N_{ij}} \times \frac{(y_i - \mu_j)}{\sigma_j^2} + C}{\sum_{i \in \{1...k\}} \frac{\delta f(\{N_{ij}\})}{\delta N_{ij}} \times \frac{(y_i - \mu_j)}{\sigma_j^2} \mu_j + D_j C} \tag{80}$$

$$\hat{\sigma}_j = E_j \frac{\sum_{i \in \{1...k\}} \frac{\delta f(\{N_{ij}\})}{\delta N_{ij}} N_{ij} [-1 + \frac{(y_i - \mu_j)^2}{\sigma_j^2}] + C\sigma_j}{\sum_{i \in \{1...k\}} \frac{\delta f(\{N_{ij}\})}{\delta N_{ij}} N_{ij} [-1 + \frac{(y_i - \mu_j)^2}{\sigma_j^2}] + E_j C} \tag{81}$$

In the case that $\mu_j$ are negative one can change coordinates to make them positive (i.e. add to $\mu_j$ some positive number), compute updates for new variables in the new coordinate system and then go back to the old system of coordinates. Q.E.D.

# 10 Comparison of two growth transformation

Here we compare two different growth transformation that were obtained in this paper for very large $C$. From the linearization principle it follows that one can consider a linear form. Here we assume that $0 < \mu < 1$ (to apply discrete transformation formula to continuous parameters).

$$l = \sum a_i N_i \tag{82}$$

$$\frac{\delta l}{\delta \mu} = \sum a_i N_i \frac{y_i - \mu}{\sigma^2} \tag{83}$$

$$\begin{aligned} l(\mu\prime) - l(\mu) &\sim \frac{\delta l_N}{\delta \mu}(\mu\prime - \mu) = \\ &= \sum a_i N_i \frac{(y_i - \mu)}{\sigma^2}(\mu\prime - \mu) = \\ &= \frac{[\sum a_i N_i \frac{y_i - \mu}{\sigma^2}]^2 \mu(1-\mu)}{\sum a_i N_i + C} = \\ &= \frac{[\sum c_i \frac{y_i - \mu}{\sigma^2}]^2 \mu(1-\mu)}{\sum a_i N_i + C} \end{aligned} \tag{84}$$

At the same time the growth using a different formulae can be expressed as

$$l(\mu\prime) - l(\mu) \sim \frac{1}{C\sigma^2}\{\frac{\{\sum c_i[(y_i-\mu)^2-\sigma^2]\}^2}{2\sigma^2} + \frac{[\sum c_i(y_i-\mu]^2}{C\sigma^2}\} \tag{85}$$

It is easily to construct examples when some of the formula provides bigger incremental step.. For example, if $\sum c_i[(y_i-\mu)^2-\sigma^2]$ is close to zero and $\sigma$ close to zero than the first formula (modified Baum) provide bigger incremental step than standard Baum.

# 11 Preliminary numerical simulation experiments

Our preliminary experiments are done for linear forms of Gaussians $l(\mu,\sigma) := l(\{N_i\}) := \sum_{i=1}^{i=m} a_i N_i$. We are taking weighted sum for two variants for growth for $\mu,\sigma$ - modified and standard Baums. We also vary constant $C$ in modified and standard Baum transformations to find for which weights and constant $C$ we have the biggest incremental step. Here are more details about the experiments.

1. Compute *best* standard Baum We compute standard Baum for linear forms for $C$ varying from $t_1$ to $t_2$

$$\hat{\mu}(C) = \frac{\sum_{\text{J}=1}^{j=m} c_j y_j + C\mu}{\sum_{\text{J}=1}^{j=m} c_j + C} \tag{86}$$

where $c_j = a_j N_j$.

$$\hat{\sigma}(C)^2 = \frac{\sum_{\text{J}=1}^{j=m} c_j y_j^2 + C(\mu^2+\sigma^2)}{\sum_{\text{J}=1}^{j=m} c_j + C} - \hat{\mu}(C)^2 \tag{87}$$

And set $\mu_s = \hat{\mu}(C\prime)$, $\sigma_s = \hat{\sigma}(C\prime)$ where optimal $C\prime = argmax_{C\in\{t_1,\ldots,t_2\}} l(\hat{\mu}(C\prime),\hat{\sigma}(C\prime))$

2. Compute *best* modified Baum We compute modified Baum for linear forms for $C$ varying from $t_1$ to $t_2$

$$\hat{\mu}(C) = D\mu\frac{\sum_{i\in\{1\ldots k\}} c_i \times \frac{(y_i-\mu)}{\sigma_j^2} + C}{\sum_{i\in\{1\ldots k\}} c_i \times \frac{(y_i-\mu)}{\sigma_j^2}\mu + DC} \tag{88}$$

$$\hat{\sigma}(C) = E\frac{\sum_{i\in\{1\ldots k\}} c_i[-1+\frac{(y_i-\mu)^2}{\sigma^2}] + C\sigma}{\sum_{i\in\{1\ldots k\}} c_i[-1+\frac{(y_i-\mu)^2}{\sigma_j^2}] + EC} \tag{89}$$

And set $\mu_m = \hat{\mu}(C\prime)$, $\sigma_m = \hat{\sigma}(C\prime)$ where optimal $C\prime = argmax_{C\in\{t_1,\ldots,t_2\}} l(\hat{\mu}(C\prime),\hat{\sigma}(C\prime))$

3. Compute *best* mixture of standard and modified Baum We define a mixture of Baums as:

$$\mu(\alpha) = \alpha\mu_s + (1-\alpha)\mu_m \tag{90}$$

$$\sigma(\alpha\prime) = \alpha\prime\sigma_s + (1-\alpha\prime)\sigma_m \tag{91}$$

And set $\hat{\mu} = \mu(\hat{\alpha})$, $\hat{\sigma} = \sigma(\hat{\alpha}\prime)$ where optimal $(\hat{\mu},\hat{\sigma}) = argmax_{(\alpha,\alpha\prime)\in[0,1]\times[0,1]} l(\mu(\alpha),\sigma(\alpha\prime))$

*Typical numerical example* Here are some experimental results along the lines that are described above. In these experiments D=E=3 in (88) and (89), a number of observables $y_i$ and coefficients $a_i$ in the linear form $l(N_i)$ equals $m$. Coefficients in this linear form $a_i$ and observables $y_i$ are random.

In the table below $\alpha_\mu = \alpha$ and $\alpha_\sigma = \alpha\prime$ from (90) and (91), Mod Baum (best C) stands for best modified Baum , Stdn Baum (with best C) stands for best standard Baum and Mix mod-stnd Baum denotes a best mixture of standard and modified Baum. Mod Baum and Stdn Baum were computed either from initial $\mu, \sigma$ or from those $\mu, \sigma$ that were obtained in a previous iteration for Mix mod-stnd Baum.

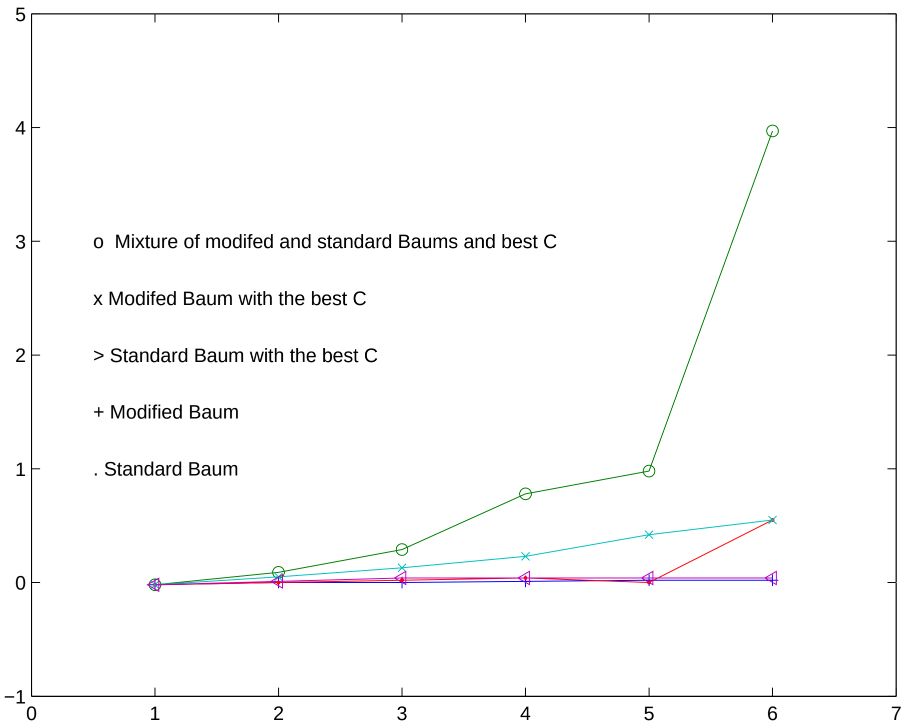


Figure 1: Graphs of objective values for 5 maximization methods .

| Iter | Method of Maximization | $\alpha_\mu$ | $\alpha_\sigma$ | C | Obj Value |
|---|---|---|---|---|---|
| 0 | —— | ——— | – | – | -0.015 |
| 1 | Mod Baum (best C) | – | – | 1.0 | 0.052 |
| 1 | Stnd Baum (best C) | – | – | 6 .0 | 0.01 |
| 1 | Mix mod-stnd Baum | 1 | 0 | – | 0.087 |
| 2 | Mod Baum (best C) | – | – | 1.0 | 0.052 |
| 2 | Stnd Baum (best C) | – | – | 11 | 0.141 |
| 2 | Mix mod-stnd Baum | 1 | 0 | – | 0.292 |
| 3 | Mod Baum (best C) | – | – | 6.0 | 0.344 |
| 3 | Stnd Baum (best C) | – | – | 66 | 0.57 |
| 3 | Mix mod-stnd Baum | 1 | 0 | – | 0.778 |
| 4 | Mod Baum (best C) | – | – | 51 | 0.97 |
| 4 | Stnd Baum (best C) | – | – | 51 | 0.778 |
| 4 | Mix mod-stnd Baum | 4/5 | 1 | – | 0.98 |
| 5 | Mod Baum (best C) | – | – | 11 | 3.96 |
| 5 | Stnd Baum (best C) | – | – | 11 | 0.981 |
| 5 | Mix mod-stnd Baum | 1/10 | 1 | – | 3.97 |

These illustrative simple numerical experiments 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 to estimating weights and constants is to treat them as parameters and estimate them together with means and variables. For example, assuming that $1 \leq C \leq \infty$ one can replace in (1) $C = 1/P$ and obtain the new formula

$$\hat{z}_{ij} = \frac{z_{ij}(P\frac{\delta}{\delta z_{ij}}R(z)+1)}{\sum_i z_{ij}(P\frac{\delta}{\delta z_{ij}}R(z)+1)} \tag{92}$$

Substituting these formula for $P$ into $R$ one can estimate $P$ using (1) transformation for $P$ (adding a slack variable $P\prime$ and constraints $P + P\prime = 1$). We will investigate this approach somewhere.

# 12 Appendix

Here we show that $C(\{a_{ij}\}) \to \infty$ when $h \to 0$.
The formulae in (1) is obtained as follows.
Let us consider $P = P_1(x) - kP_2(x) + C'f(x)$ where $f(x)$ is the constant over a domain of probability values, $k = \frac{P_1(x_0)}{P_2(x_0)}$, $C'$ is such a large constant that $P$ has positive coefficients. Then a growing transformation for the polynomial $P$ is defined as follows:
$\hat{x}_{ij} = \frac{x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')}{\sum_i x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')}$
$\hat{x}_{ij} = \frac{x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')}{\sum_i x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')}$ Where $C'' = \frac{\delta}{\delta x_{ij}}C'f(x)$ is the constant (independent of $i, j$). The formulae (1) can be obtained as follows:

$$\begin{aligned}\hat{x}_{ij} &= \frac{x_{ij}(\frac{\delta}{\delta x_{ij}}(P_1(x)-kP_2(x))+C'')}{\sum_i x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')} = \frac{x_{ij}(\frac{\delta}{\delta x_{ij}}P_1(x) - \frac{P_1(x_0)}{P_2(x_0)}\frac{\delta}{\delta x_{ij}}P_2(x)+C'')}{\sum_i x_{ij}(\frac{\delta}{\delta x_{ij}}P(x)+C'')} \\ &= \frac{x_{ij}(P_2(x_0)\frac{\delta}{\delta x_{ij}}P_1(x)\frac{1}{P_2^2(x_0)} - \frac{P_1(x_0)}{P_2^2(x_0)}\frac{\delta}{\delta x_{ij}}P_2(x) + \frac{C''}{P_2(x_0)})}{\sum_i x_{ij}(\frac{\delta}{\delta x_{ij}}\frac{P(x)}{P_2(x_0)} + \frac{C''}{P_2(x_0)})}\end{aligned} \tag{93}$$

For $x = x_0$ we get (1) with $C = \frac{C''}{P_2(x_0)}$. If $P_2$ is a homogenous polynomial and all coordinates $x = (x_{ij}) \to 0$ then $C \to \infty$. This is the case when $x_{ij} = a_{ij}$ in (4). Namely, $x_{ij} = a_{ij} \to 0$ if $h \to 0$.

# References


[1] Scott Axelrod, Vaibhava Goel, Ramesh Gopinath, Peder Olsen, and Karthik Visweswariah, "Discriminative Training of Subspace Constrained GMMs for Speech Recognition," to be submitted to IEEE Transactions on Speech and Audio Processing.

[2] Scott Axelrod, Vaibhava Goel, Ramesh Gopinath, Peder Olsen, and Karthik Visweswariah, "Personal communication."

[3] 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.

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

[5] Ponani S Gopalakrishnan , Dimitri Kanevsky, David Nahamoo, A. Nadas *An inequality for rational functions with applications to some statistical estimation problems*, IEEE Trans. Information Theory, Vol. 37, No.1 January 1991

[6] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo, *Maximum Mutual Information Training of Hidden Markov Models with Continuous Parameters*, IBM Technical Disclosure Bulletin, Vol. 33. , Dec 1990

[7] Kanevsky, D., *A generalization of the Baum algorithm to functions on non-linear manifolds*, In: Proc. Internat. Conf. On Acoustics, Speech and Signal Processing, May 1995, Detorit, MI, Vol. 1., pp.473-476.

[8] Y. Normandin (1991), *Hidden Markov Models, Maximum Mutual Information Estimation, and the Speech Recognition Problem*, Thesis, March 1991, McGill University, Montreal.

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