A GENERALIZATION OF THE BAUM ALGORITHM TO FUNCTIONS ON NON-LINEAR MANIFOLDS

D. Kanevsky

Human Language Technologies,
IBM T.J. Watson Research Center,
P.O. Box 704 , Yorktown Heights N.Y., 10598

ABSTRACT

The well-known Baum-Eagon inequality [1] provides an effective iterative scheme for homogeneous polynomials with positive coefficients over a domain of probability values Δ. In [2] the Baum-Eagon inequality was extended to rational functions over Δ and in [3,4] a variant of this extended inequality was used for the maximum mutual information training of a connected digit recognizer.
However, in many applications (e.g. corrective training) we are interested in maximizing an objective function over a domain D that is different from Δ and may be defined by non-linear constraints. In the paper we show how to extend the basic inequality from [2] to (not necessary rational) functions that are defined on general manifolds. We describe an effective iterative scheme that is based on this inequality and its application to estimation problems via minimum information discrimination.

1. Introduction

Let us formulate a problem that we consider in the paper as follows.

Problem: Let Δ be a domain of probability values:

$$(1) \qquad \Delta: \quad x_{ij} \geq 0, \quad \sum_{j=1}^{t_i} x_{ij} = 1, \quad i = 1, \dots, r$$

Let $D \subset \Delta$ be a sub-manifold and f be a (single-valued) function on D. We want to find a growing transformation $T: D \to D$, i.e. such that $f(T(x)) \geq f(x)$ for any $x \in D$.
Our primary interest here is in problems that involve a large number ($n =$ several thousands) of variables (x_{ij}) and therefore optimization methods that require calculation of the Hessian matrix would be computationally infeasible. For example, standard implementation of a Newton's method requires too much storage ($O(n^2)$) and too much work per iteration ($O(n^3)$ flops).

2. Main Results

In order to formulate our new inequality we need to introduce the following definition:

Projection: Let $f(x)$ be differentiable at $x = \{x_{ij}\} \in D$. Let $C = \{C_1, \dots, C_r\}$ be a set of non-negative constants, $\xi_{ij}^C(x) = x_{ij} \frac{\partial f(x)}{\partial x_{ij}} + C_i x_{ij}$ and $\xi_i^C(x) = \sum_{j=1}^{t_i} \xi_{ij}^C(x) \neq 0$ for all i (we set $\xi_{ij}(x) = \xi_{ij}^C(x)$ and $\xi_i(x) = \xi_i^C(x)$ if $C = 0$).

Then a 'projection' $T^C = T^{f,C}$ at x is defined as follows:

$$(2) \qquad T^C(x)_{ij} = \frac{\xi_{ij}^C(x)}{\xi_i^C(x)}$$

where $T^C(x)_{ij}$ denotes the $ij-$ coordinate of the vector $T^C(x)$.

The following theorem states that under very general assumptions on a manifold D T^C is a growing transformation for a sufficiently large C.

Theorem 1: Let $x \in D \subset \Delta$ be an inner point of D, i.e. there is an open ball $U(x) = \{y: |x - y| < \varepsilon\}$ (where $|...|$ denotes the Euclidean distance) such that $U(x) \subset D$. Let f be analytic at x (i.e. f can be represented locally by power series). Then there exists $C \geq 0$ such that for all C with $C_i \geq C_0$ the following holds: $T^C(x) \in D$ and $f(T^C(x)) \geq f(x)$.

Proof: In order to prove the theorem we need to approximate the function $f(x)$ by polynomials for which the statement of the theorem holds. After this the general statement is preserved under limit transition. In what follows we describe these steps in greater detail.
Step 1 (Uniform approximation): Let $f(x)$ is represented as a power series $F(x)$ in some neighborhood of x and let a polynomial $f_m(x)$ consists of all terms of $F(x)$ of degree less than m. Then $f_m(x)$ converges *uniformly* in

0-7803-2431-5/95 $4.00

some neighborhood of x. Let $h_m(x) = f_m(x) + C_m(x)$, where $C_m(x) = -C_m \times (\sum_{i=1,j=1}^{r,t_i} x_{ij} + 1)^m$ and C_m equals

a minimal negative coefficient in f_m (or zero if there are no negative terms). Let $f_m = h_m(x) - D_m$ where the constant $D_m = C_m(x)$ on D. Then all non-zero-degree coefficients of $f_m(x)$ are nonnegative and $f_m(x)$ approximates $f(x)$ *uniformly* in some neighborhood U of x. In other words, for any $\varepsilon \geq 0$ there exists an integer N such that for all $m \geq N$ and $y \in U$ $|f(y) - f_m(y)| \leq \varepsilon$.

Step 2 (Inequality for polynomials): The inequality in [1] is formulated for polynomials with nonnegative coefficients. But it is easily to see that zero-degree coefficients do not affect Baum-Eagon transformation formulas. Therefore the following inequality holds: $f_m(T_m^C(x)) \geq f_m(x)$ for any $C \geq 0$, where T_m^C is the projection (2) for $f_m(x)$. When $C \to \infty$ $T_m^C(x) \to x$ and therefore there exists C_m such that for all $C \geq C_m$ $T_m^C(x) \in U$. In other words the theorem 1 holds for polynomials $f_m(x)$.

Step 3 (Carrying over to limits): Since $\xi_i(x) \neq 0$ for sufficiently large m and C the function $T_m^C(x)$ is defined at x and $T_m^C(x) \to T^C(x)$. This implies that there exists C_x such that for all $C \geq C_x$ $T_m^C(x) \in U$ for sufficiently large m. Carrying m to infinity and using inequalities from Step 2 gives the full statement of Theorem 1.

In practical applications it is useful to have the following equivalent of Theorem 1.

Theorem 2: Let conditions of Theorem 1 are fulfilled. Let $\beta_i = 1 - \alpha_i$, $B(x)_{ij} = T^v(x)_{ij}$ (with $C = 0$ in (2)) if $\xi_i(x) > 0$ and $\beta_i = 1 + \alpha_i$, $B(x)_{ij} = -T^v(x)_{ij}$ otherwise. Let

$$x(\alpha)_{ij} = B(x)_{ij} \times \alpha_i + x_{ij} \times \beta_i \tag{3}$$

Then there exists ε such that for all $0 \leq \alpha < \varepsilon$ the following holds: $x(\alpha) \in D$ and $f(x(\alpha)) \geq f(x)$.

The fact that these both theorems are equivalent is easily follows from the following observation.

Lemma: In conditions of theorems 1,2 the following holds:

$$T^C(x) = T^v(x)_{ij} \times \alpha'_i + x_{ij} \times (1 - \alpha'_i) \text{ where } \alpha'_i = \frac{\xi_i(x)}{\xi_i^C}.$$

In particular, $\alpha' = 1/C + o(1/C)$ if $\xi_i(x) > 0$ and $\alpha' = -1/C + o(1/C)$ otherwise. (Here $C \times o(1/C) \to 0$ when $C \to \infty$).

3. Comments

1. For sufficiently small α the following estimate of the function growth holds:

$$f(x(\alpha)) - f(x) \geq \sum_{i=1}^{r} l_i(x) \times \alpha + o(\alpha) \tag{4}$$

where $l_i(x) = (\frac{\hat{\xi}_i(x)}{\xi_i(x)} - \xi_i(x))$, $\hat{\xi}_i = \sum_{j=1}^{t_i} x_i \frac{\partial f(x)^2}{\partial x_{ij}}$ and

$o(\alpha)/\alpha \to 0$ when $\alpha \to 0$.

2. Using estimate (4) one can show that in above theorems $f(x(\alpha)) > f(x)$ if $T^v(x) \neq x$ and $\alpha \neq 0$. Also it is easily to show that if $T^v(x) = x$ then standard necessary first-order maximum conditions with Lagrange multipliers at x are satisfied and therefore $f(x)$ has a local maximum at x if $f(x)$ is concave in a neighborhood of x.

3. One can show that theorems 1,2 remain true under an weaker condition that $f(x)$ has first-order derivatives in some open neighborhood of x.

4. In the maximum likelihood training when an objective function $f(x)$ is a polynomial with non-negative coefficients we have: $f(x(\alpha)) \geq f(x)$ for any $0 \leq \alpha \leq 1$.

5. Comments 1,2 imply the following statement. Let $D_x = \{\alpha | x(\alpha) \in D\}$. Let $\tilde{\alpha}$ is a set of the 'best' step sizes along $T(x)$, i.e.

$$\tilde{\alpha} = \{\tilde{\alpha}_i\} = \arg\max_{\alpha \in D_x} f(x(\alpha)) \tag{5}$$

Then $f(x(\tilde{\alpha})) > x$ if $T^v(x) \neq x$.
The problem (5) is easier than the original problem since it involves less number of parameters than in (1).

6. In practical applications it is useful to have transformation formulas for the following (slightly) more general domain:

$$\Delta: \quad x_{ij} \geq 0, \quad \sum_{j=1}^{t_i} x_{ij} = a_i > 0, \quad i = 1, \ldots, r \tag{6}$$

In this case the growth transformation like in (2) can be defined as follows:

$$T^C(x)_{ij} = a_i \times \frac{\xi_{ij}^C(x)}{\xi_i^C(x)} \tag{7}$$

4. Variations

In [3] Merialdo considered the problem of maximizing a rational function (mutual information) over a domain of probability values and suggested to separate positive and negative terms $\xi_{ij}(x)$ and to form a new gradient as the difference of projections of these positive and negative terms. In [4] Normadin et al. suggested to add a large constant to the Merialdo gradient (in the spirit of [2]). This work inspired the following straightforward generalization of Theorem 2.

Modified projection: Let $f(x)$ be differentiable at $x = \{x_{ij}\} \in D$ and let $\xi_i(x) \neq 0$ for all i. Let $S_+(i) = \{j | \xi_{ij}(x) > 0\}$ be a subset of indexes for which $\xi_{ij}(x)$ is positive and let $S_-(i) = \{j | \xi_{ij}(x) < 0\}$ complements $S_+(i)$. Let $\xi_i^+(x) = \sum_{j \in S_+(i)} \xi_{ij}(x)$. Similarly, let $\xi_i^-(x) = \sum_{j \in S_-(i)} \xi_{ij}(x)$.

Let $a_i^+ = \sum_{j \in S_+(i)} x_{ij}$ and let $a_i^- = \sum_{j \in S_-(i)} x_{ij}$.

Let

$$(8) \qquad T_f^+(x)_{ij} = a_i^+ \times \xi_{ij}(x) / \xi_i^+(x)$$

for $i \in S_+(i)$, $T_f^+(x)_{ij} = 0$ *for* $i \in S_-(i)$
And let

$$(8') \qquad T_f^-(x)_{ij} = a_i^- \times \xi_{ij}(x) / \xi_i^-(x)$$

for $i \in S_-(i)$, $T_f^-(x)_{ij} = 0$ *for* $i \in S_+(i)$.
Then, finally, 'modified projection' is defined as follows:

$$(9) \quad T_\alpha(x)_{ij} = T_f^+(x)_{ij} \times \alpha_i^+ - T_f^-(x)_{ij} \times \alpha_i^- + x_{ij} \times \beta_{ij}$$

where $\alpha = \{\alpha_i^+, \alpha_i^-\}, i = 1, \ldots, r$ and $\beta_{ij} = 1 - \alpha_i^+$ if $j \in S_+(i)$ and $\beta_{ij} = 1 + \alpha_i^-$ if $j \in S_-(i)$.

This projection defines a growing transformation for sufficiently small α. The following statement follows immediately from the comment 6 in Chapter 3.

Theorem 3: Let conditions of Theorem 1 are fulfilled. Then there exists $\varepsilon > 0$ such that for all $0 \leq \alpha_i^+, \alpha_i^- < \varepsilon$ $T_\alpha(x) \in D$ and $f(T_\alpha(x)) \geq f(x)$.

Remarks: Formulas (8),(8') allow to introduce new parameters that control the optimization procedure and therefore can be useful to adjust gradient to constraints when x approaches the boundary of D. There exist also modifications of (8),(8') that allow to improve convergence rate in an iterative optimization procedure (see Numerical Experiments in Section 5).

5. Application to Minimum Information Discrimination

In this section we apply this theorem to the I-divergence function: $f(x) = I(x||q) = \sum_{i=1}^{n} x_i \log(\frac{x_i}{q_i})$ (where $x, q \in \Delta$, $r = 1, t_i = n$ in (1) and $x \neq q$), that plays an important role in statistics (see [5]). The formula (3) gives rise to the following 'decreasing' transformation:

$$(10) \qquad T(x)_i = -\alpha \frac{I_i(x||q) - Cx_i}{I(x||q) - C} + \beta x_i$$

where $I_i(x||q) = x_i \log(\frac{x_i}{q_i})$, and $\beta = 1 + \alpha$. (The minus sign in the formula (10) appears since we are minimizing f(x), and α and β are chosen in such a way that the sum of the components of the vector $T(x)$ in (3) equals 1). This transformation has the property that for sufficiently small α $I(T(x)||q) \leq I(x||q)$ and $T(x) \in D$ if $x \in D$ is an inner point.
Similar formulas can be produced using (8),(8').

Using the decreasing transformation (10) one can suggest the following sub-optimal scheme for minimizing the discrimination function $f(x) = I(x||q)$ over a manifold $D = \{x \in \Delta : f_j(x) > 0, j = 1, \ldots, m\}$, where f_j are linear or quadratic functions of n variables x_i.

Iterative scheme : 0) Start with some point $x_0 \in D$, $t = 0$ and small α_0; 1) x^{t+1} is computed by (10) (or (8), (8')) with α_t chosen sufficiently small to satisfy linear and quadratic inequalities $f_j(x^{t+1}) > 0$ (one can use 5. in Comments for funding suboptimal α_t); 2) $t = t + 1$, go to 1).
Since $I(x||q)$ is convex, one can show that x^t converges to some point on the boundary of D if q does not belong to D.

Note that in general a domain D does not need to be convex and therefore this iterative scheme can be especially useful when standard methods for minimization over convex domains (e.g. [5]) are not applicable.

Remarks: Useful models could be created by minimizing $I(x||q)$ subject to the following constraints: $x \in \Delta$, $\sum(x_i - \tilde{q}_i)^2 < \varepsilon$, where sumation is taken over some subset of indexes and $\tilde{q}$ are some observed frequences. For example, in the language model training when q represent unigram, bigram or trigram probabilities, $\tilde{q}$ could be a set of frequences that were derived from additional textual data (e.g. key words, topics etc.). Note that when we are reaching a point on a bound using the above training scheme (in other words we get a point x satisfying the quadratic equation $\sum(x_i - \tilde{q}_i)^2 = \varepsilon$) one can continue the training procedure

using some version of the formula (9) and choosing weight coefficients α in such a way that the above quadratic equality is fulfilled. We'll report elsewhere about this kind of training.

Numerical experiments: We run experiments using formulas (10) for different values of α to see how the decrease of Kullback Distance $I(x||q)$ could be controlled. We set $n = 1{,}000$, $D = \Delta$ and chose randomly q and starting points x in D. We also set in (10) $C = \max\{log(x_i/q_i)\} + \varepsilon$ (where ε was some fixed small constant) and varied α between 0 and 1. This choice of C guarantees that $T(x)$ in (10) belongs to Δ (for any $0 < \alpha \leq 1$). The table 1. shows values of $I(x'(\alpha)||q)$ each 5 iterations (until the 20th iteration) for fixed $\alpha = 1$ and 1/3 (for a typical example). The table shows that the fastest decrease of $I(x'(\alpha)||q)$ was for the first 5 iterations for $\alpha = 1$ and lower values of Kulback Distance were achieved for $\alpha = 1/3$ after $t = 15$. For $\alpha = 1$ the value of $I(x'(\alpha)||q)$ did not decrease gradually after the 6th iteration but rather circulated between 0.14 and 0.068 with each iteration. For $\alpha = 1/3$ the value of $I(x||q)$ decreased gradually until the 19th iteration. This experiment suggests , as it was expected, that should start optimize the I-divergence function using some values of α near 1 and then gradually decrease α while approaching a minimum point. In practical applications choice of α would also be affected by additional imposed constraints.

In the other experiment (following remarks in Section 4) we studied transformations:

$$(11) \qquad T(x)_i = \alpha \times (T_1(x)_i - T_2(x)_i) + \beta \times x_i$$

Here: $T_2(x)_i = \dfrac{I_i(x||q)}{I_+(x||q)}$ if $x_i > q_i$ and 0 otherwise;

$T_1(x)_i = \dfrac{I_i(x||q)}{I_-(x||q)}$ if $x_i < q_i$ and 0 otherwise;

$I_+(x||q) = \sum_{(i:x_i > q_i)} I_i(x||q)$, $I_-(x||q) = \sum_{(i:x_i < q_i)} I_i(x||q)$.

In other words, we separate terms in $I(x||q)$ in accordance with the sign of $\log(x_i||q_i)$.
In the table 2. we gave values of $I(x||q)$ for the first 4 iterations using transformations (11) with $\alpha = 1/10$. After the 4th iteration the value of $I(x'(\alpha)||q)$ circulated between 0.014 and .006 .

6. Summary

In the paper we generalized the Baum-Eagon inequality to general functions on non-linear manifolds. Following this general theory we calculated transformation formulas for the I-divergence function. We produced numerical data that demonstrate that these transformations effectively decrease I-divergence. We also provided numerical evidence that separating positive and negative terms in a gradient one can improve convergence rate in an iterative optimization procedure.

Acknowledgment

We wish to thank Patibanda Srinivasa for many useful discussions.

References

1. 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.
2. P.S. Gopalakrishnan, D. Kanevsky, A. Nadas and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems", *IEEE Trans. Inform. Theory*, Jan. 1991.
3. B. Merialdo, "Phonetic recognition using hidden Markov models and maximum mutual information training", *Proc. ICASSP-88*, 1988. *IEEE Trans. Acoust., Speech, Signal Processing* , April 1994.
4. Y. Normadin, R. Cardin and R. De Mori, "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation", *IEEE Trans. Acoust., Speech, Signal Processing* , April 1994.
5. I. Csiszar, "I-divergence geometry of probability distributions and minimization problems", *Ann. Probab.* Vol. 3, pp. 146-158.

Table 1. Transformations (10)

$I(x'(\alpha)\|\|q)$	$t = 0$	$t = 5$	$t = 10$	$t = 15$	$t = 20$
$\alpha = 1$	0.4818	0.0444	0.1495	0.0675	0.1400
$\alpha = 1/3$	0.4818	0.2284	0.0855	0.0075	0.0001

Table 2. Transformations (11)

$I(x'(\alpha)\|\|q)$	$t = 0$	$t = 1$	$t = 2$	$t = 3$	$t = 4$
$\alpha = 1/10$	0.4818	0.2544	0.1011	0.0185	0.0059