Convergence of Line Search A-Function methods

Dimitri Kanevsky, David Nahamoo, Tara Sainath, Bhuvana Ramabhadran

IBM T.J. Watson Research Center, Yorktown, NY 10598, USA
{kanevsky, nahamoo, tsainath, bhuvana}@us.ibm.com

Abstract
Recently, the Line Search A-Function (LSAF) was introduced as a technique that generalizes Extended Baum-Welch (EBW) algorithm for functions of continuous probability densities. It was shown that LSAF provides a unified scheme for a large class of optimization problems that involve discriminant objective functions of different probability densities or sparse representation objective functions such as Approximate Bayesian Compressive Sensing. In this paper, we show that a discrete EBW recursion (that was initially developed to optimize functions of discrete distributions) also fits the scope of LSAF technique. We demonstrate the utility and robustness of the technique for discrete distributions through the experimental set up of a TIMIT phone classification task using a Convex Hull Sparse Representation approach with different Lq regularization (q being any positive number).

Index Terms: EBW, convergence, A-function, LSAF

1. Introduction
The Extended Baum-Welch (EBW) technique was initially introduced for estimating the discrete probability parameters of multinomial distribution functions of HMM speech recognition problems under the Maximum Mutual Information discriminative objective function [1]. Later, in [9], EBW technique was extended to estimating the parameter of Gaussian Mixture Models (GMMs) of HMMs under the MMI discriminative function for speech recognition problems. In ([5]) the EBW technique was generalized to the novel Line Search A-functions (LSAF) optimization technique. In ([5]), a simple geometric proof was provided to show that LSAF recursions result in a growth transformation (i.e. the value of the original function increases for the new parameters values). In the paper we show that a discrete EBW that was invented more than 23 years ago can be also represented using A-functions (like its continuous EBW version). We also for the first time give a convergence proof for a discrete EBW. This technique of proof can be easily applied also for proving convergence of continuous EBW (that optimize functions of GMMs) and LSAF process. Finally, we provide experimental results for TIMIT phonetic classification task using fractional norm generalization of a recently introduced Convex-Hull sparse representation technique that deploys discrete EBW for optimization ([10]). The rest of the paper is structured as follows. In Section 2, we provide a short review of the LSAF method which was introduced in ([5]). In Section 3 we show that a discrete EBW can be represented in the LSAF framework. In Section 4 we outline a proof of convergence of a discrete EBW. In Section 5 we describe a discrete EBW recursion for optimization problems with fractional norms constraints. In Section 6 we demonstrate this techniques for phonetic tasks ([8]). In Section 7 we give conclusions and in Appendix we give a detailed proof of convergence for EBW.

2. LSAF
2.1. Definition
Let \( f(x) : U \subset \mathbb{R}^n \to \mathbb{R} \) be a real valued differentiable function in an open subset \( U \). Let \( A_f = A_f(x,y) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \) be twice differentiable in \( x \in U \) for each \( y \in U \). We define \( A_f \) as an \( A \)-function for \( f \) if the following properties hold.

1. \( A_f(x,y) \) is a strictly convex or strictly concave function of \( x \) for any \( y \in U \) (recall that twice differentiable function is strictly concave or convex over some domain if its Hessian function is positive or negative definite in the domain, respectively).
2. Hyperplanes tangent to manifolds defined by \( z = g_f(x) = A_f(x,y) \) and \( z = f(x) \) at any \( x = y \in U \) are parallel to each other, i.e. \( \nabla_z A_f(x,y) |_{y=x} = \nabla_z f(x) \).

It was shown in ([5]) that a general optimization technique can be constructed based on \( A \)-function. We formulated a growth transformation such that the next step in the parameter update that increases \( f(x) \) is obtained as a linear combination of the current parameter values and the value that optimizes the \( A \)-function, i.e. \( \nabla_z A_f(x,y) |_{y=x} = 0 \). More precisely, we stated that \( A \)-function gives a set of iterative update rules with a "growth" property as the following. Let \( x_0 \) be some point in \( U \) and \( U \ni x_0 \neq x_0 \) be a solution of \( \nabla_z A_f(x,y) |_{y=x_0} = 0 \) (it is the minimum of \( A_f(x,y) \) if \( A_f \) is concave and the maximum if \( A_f \) is convex). Let \( x_1 = x_0 - \alpha x_0 + (1 - \alpha)x_0 \).

Then for sufficiently small \( \alpha \neq 0 \), \( f(x_1) > f(x_0) \) where \( \alpha > 0 \) if \( A(x_0, x_0) \) concave and \( \alpha < 0 \) if \( A(x_0, x_0) \) convex.

We called this technique Line Search A-Function (LSAF).

2.2. Gaussian example
Let \( \xi(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\} \), be densities over a space \( x \in \mathcal{X} \) with model parameters \( \theta = \{\mu, \sigma\} \) and let us consider a function \( f((\xi(\theta))) \). Let \( c_\theta(\theta) = \frac{\theta}{\sigma^2} \). Then

\[ \text{Assoc}(\xi(\theta_0), \xi(\theta)) = \sum c_\theta(\theta_0) \log f((\xi(\theta))) \]

is \( A \)-function. This gives us the following updates for model parameters:

\[ \bar{\mu} = \bar{\mu}(\alpha) = \alpha \bar{\mu}_0 + (1 - \alpha) \mu_0, \quad \bar{\sigma}^2 = \bar{\sigma}^2(\alpha) = \alpha \bar{\sigma}_0^2 + (1 - \alpha) \sigma_0^2 \]

where \( \bar{\mu}_0 = \frac{1}{z_0(c_\theta(\theta_0))} \) and \( \bar{\sigma}_0^2 = \frac{1}{z_0(c_\theta(\theta_0))} \). Since \( \text{Assoc}(\xi(\theta_0), \xi(\theta)) \) is an \( A \)-function, updates (3) are growth for sufficiently small \( |\alpha| \). It was shown in ([5]) that EBW recursions for functions of Gaussian densities belong to this family of transformations (3).
3. Discrete EBW

In this section we show that discrete EBW can be described using the LSAF framework. In what follow we describe the case of a single distribution but this technique is easily generalizable to several distributions.

Let \( S = \{ \beta : \beta \in \mathbb{R}^n, \beta_i \geq 0, i = 1, \ldots, n, \sum \beta_i = 1 \} \) and \( f : \mathbb{R}^n \to \mathbb{R} \) be a differentiable function on some subset \( X \subset S \). We want to solve the following maximization problem for a function \( f(\beta) \)

Maximize \( f(\beta) \) such that \( \beta \in S \) \hspace{1cm} (4)

Let for \( \beta \in X \) \( a_i = \frac{\partial f(\beta)}{\partial \beta_i}, i = 1, \ldots, n. \) For any \( D \in \mathbb{R} \) and \( \beta^k \in \mathbb{R}^n \) such that \( \sum_{i=1}^n a_i \beta_i^k + D \neq 0 \) we define a recursion \( T_D : \mathbb{R}^n \to \mathbb{R}^n \) as

\[
\beta_i^{k+1} = T_D(\beta^k) = \frac{a_i \beta_i^k + D \beta_i^k}{\sum_{j=1}^n a_j \beta_j^k + D}
\hspace{1cm} (5)
\]

It was shown in [2] that for sufficiently large \( D \) if \( \beta^{k+1} \neq \beta^k \) then \( f(\beta^{k+1}) > f(\beta^k) \). An A-function \( A_f \) for the function \( f \) in (4) differentiable in some compact neighborhood \( U \subset X \) of a point \( \beta_0 \in S \) is defined as following:

\[
A_f(\beta_0, \beta) = \sum (c_i + \beta_0 i) \log \beta_i
\hspace{1cm} (6)
\]

where \( c_i = c_i(\beta_0) = \beta_0 \frac{\partial f(\beta_0)}{\partial \beta_i} \) and \( D \) is any number such that \( a_i(\beta) + D > 0 \) for all \( i \) and any \( \beta \in U \) (such \( D \) can be found since \( f \) is differentiable in \( U \) and \( U \) is compact). To show that the function \( A_f(\beta_0, \beta) \) in (6) is a function one need to check (1) as following. Replace \( \beta_i = 1 - \sum \beta_j \) in (4, 6), i.e. consider functions \( g(\beta') = f(\beta_1, \ldots, \beta_{n-1}, 1 - \sum_{j=1}^{n-1} \beta_j) \), \( A_g(\beta_0, \beta') = A_f(\beta_0, \beta') = \sum (c_i + \beta_0 i) \log \beta_i \) where \( \beta' = \{ \beta_1, \ldots, \beta_{n-1} \} \). We have:

\[
\frac{\partial A_g(\beta_0, \beta')}{\partial \beta_i} \mid_{\beta_i = \beta_0 i} = a_i(\beta_0) \frac{\partial f(\beta_0)}{\partial \beta_i} + D \frac{\partial \log(\beta_i)}{\partial \beta_i} |_{\beta_i = \beta_0 i} + D \frac{\partial \log(1 - \sum_{j=1}^{n-1} \beta_j)}{\partial \beta_i} |_{\beta_i = \beta_0 i} = \frac{\partial g(\beta')}{\partial \beta_i} |_{\beta_i = \beta_0 i}
\]

4. Outline of convergence proof

We consider the optimization problem (4) and its EBW recursion (5). We first define the convergence statement for this problem.

For this we note that, necessary and/or sufficient conditions for some point to be a local maximum of a general constrained problem require this point to be a critical point. The notion of a critical point usually requires the derivative of the Lagrangian (that is a weighted sum of the gradient of an objective function and gradients of constraints equations) to be equal to zero as well as a system of equalities and inequalities associated with constraint equations. It can be easily seen that if a critical point of the A-function results into a fixed point in the recursion formula, this critical point will also become the critical point of the objective function \( f \). Since each A-function at each iteration is convex with a single critical point, at each iteration \( n \) we create a new \( \beta_n \) which is a critical point of the A-function at that iteration. The sequence of \( \beta_1, \beta_2, \ldots, \beta_n \) has a set of limit points since for discrete EBW the sequence of \( \beta \) belongs to a bounded set (a simplex). Based on Lemma 2 in Appendix, we know that each limit point is a fixed point with respect to the recursion formula. Depending then on regularity conditions at a critical point we get either necessary or and sufficient conditions for this point to be a local extremum.

In our special case of a discrete EBW we use the following approach. First, we notice that if there exists some entry \( \beta_r^k = 0 \) then for all consequent updates we have \( \beta_r^k = 0, r > k \). Therefore we consider a starting point \( \beta^k \) for the problem (4) such that all \( \beta_i > 0 \). And for each \( k \)-th iteration (5) we consider \( D \) so large to ensure that \( \beta^{k+1} > 0 \) for all \( i \). This means that the problem (4) has a local maximum with some entries equal zero then these entries should be obtained as entries of a limit of an infinite sequence of updates that have limit points over which \( f \) is not defined (for example, this updated sequence converge to a pole, i.e. to a point where \( f \) has the “infinite” value).

Next, to define Lagrangian for the problem (4) we consider only an equality constraint \( \sum \beta_i = 1 \) to define critical points (of a Lagrangian). We then show that the update sequence to this problem has limits points that are critical points of this constrained problem. We also show that each update in this sequence satisfies inequalities \( \beta \in S_0 = \{ \beta \in S, \beta_i > 0, i = 1, \ldots, n \} \). Therefore limits points of this sequence still belong to the domain \( S \). In Appendix we give a detailed proof following this strategy.

5. EBW method for fractional norms

We now show that discrete EBW methods can be applied to optimization of objective functions with fractional norm constraints as was suggested in ([7]):

\[
\max f(\{ \beta_i \}) \text{ s.t. } \| \beta \|_q = 1 \text{ and } \beta_i \geq 0, i = 1, \ldots, n
\hspace{1cm} (7)
\]

where \( \| \beta \|_q = (\sum \beta_i^q)^{1/q} \). Changing variables and function as in

\[
\gamma_i = \beta_i^{1/q}, g(\gamma_i) = f(\{ \beta_i \})
\hspace{1cm} (8)
\]

transforms the problem (7) into a discrete EBW problem for which recursion (5) could be applied. We will demonstrate this method for phonetic classification tasks using the objective function from ([10])

\[
F(\beta) = -1/2(\gamma - H\beta)^T R^{-1}(\gamma - H\beta) + \sum_{t=1}^{N_{classes}} \sum_{i=1}^{C_t} \sum_{i=1}^{N_t} \exp\left(-1/2 \times \left(\mu_{st} - H\beta\right)^T (\mu_{st} - H\beta) / \sigma_{st}^2\right)
\hspace{1cm} (9)
\]

Here the unique GMM classes are denoted by \( t, C_t \) is the number of Gaussians belonging to class \( t \), \( \mu_{st}, \mu_{st} \) and \( \sigma_{st} \) are the weight, mean and variance parameters for Gaussian \( i \) in GMM \( t \).
6. Experiments

Classification experiments are conducted on TIMIT acoustic phonetic corpus using feature representation that is similar to the classification work in ([8], [10]). We performed the following three experiments on functions (9) using the TIMIT dataset with constraints \( \sum \beta_i^2 = 1 \) for three values \( q = 0.5, 0.75, 1 \). We run 60 iterations of a (5) for (7) (with transformation of parameters as described in (8)). We chose \( D = 100 \) at the first iteration and changed it at each iteration by multiplying by 1.02. Dimension of \( \beta \) in our experiments was 200. For each \( q \) we chose the uniform initial \( \beta \) with entries \( 1/200^{1/q} \). Our plots for one phone (Figure 1) demonstrate that we have the robust method to update model parameters and obtain a sparse representation. We also observe that the sparsity degree (i.e. number of near zero entries) of \( \beta \) decreases with the growth of \( q \) as expected.

![Fig. 1. Likelihood functions and sparsity representation for \( \beta \) for different values of \( q \) for 60 iterations](image)

We also described in Table 1 the accuracy results on the test Timit set. The classification rule in this experiment is a combination of residual error between \( y - \mathbf{H} \beta \) and a GMM score. Number of iterations that is given in each column as iter # and \( D \) were tuned on a development set to optimize accuracy. The accuracy result for \( q = 1 \) coincides with the result that was reported in ([10]) and is the best reported number to date.

<table>
<thead>
<tr>
<th>( L_q )</th>
<th>( q=0.5/\text{iter 6} )</th>
<th>( q=0.75/\text{iter 4} )</th>
<th>( q=1.0/\text{iter 15} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>( 81.93 )</td>
<td>( 82.95 )</td>
<td>( 85.11 )</td>
</tr>
</tbody>
</table>

Table 1: Accuracy of SR method. TIMIT test set

7. Conclusions

In this paper, we showed that the discrete EBW method can be formulated in the LSAF framework. We provided a convergence proof for the discrete EBW recursion scheme. We showed how discrete EBW technique can be used to solve sparse representation problems with fractional norm constraints. By applying the technique to TIMIT phonetic classification task, we demonstrated that our method gives robust results, i.e. objective function values have a steady growth and we get a sparse representation. Finally, we believe that we can generalize our convergence proof of the discrete EBW to the broad LSAF case with the goal of presenting the generalized proof in a later paper.

8. References


9. Appendix: Convergence proof

For any map \( T \) we say that \( \beta \) is fixed if \( \mathbf{T}(\beta) = \beta \).

**Lemma 1** If \( \beta \) is fixed for \( T_D \) then \( \beta \) is a critical point of the Lagrange function

\[
A(\beta, \lambda) = f(\beta) - \lambda(\sum \beta_i - 1)
\]

for the constrained problem

\[
\text{Maximize } f(\beta) \text{ such that } \sum \beta_i = 1
\]

In addition, if the Hessian of \( f \) at \( \beta \in S_0, i \) is negative definite then \( \beta \) is a local maximum of the problem (11)

**Proof**

If

\[
\beta_i = T_D(\beta) = \frac{a_i \beta_i + D_i \beta_i}{\sum_j=1^N a_j \beta_j + D}
\]

999
then \( \beta; \sum_{j=1}^{n} a_j \beta_j = a_i \beta_i \), i.e.
\[
\sum_{j=1}^{n} a_j \beta_j = a_i
\] (13)

Next, \( \nabla_2 \Lambda(\beta, \lambda) = 0 \) implies that \( \lambda = a_i \) for all \( i \) or \( \lambda = \sum a_i \beta_i \) and this coincides with (13), i.e., with (12). The condition \( \nabla_1 \Lambda = 0 \) is exactly the constraint \( \sum \beta_i = 1 \). Next, the Hessian for the Lagrangian (10) coincides with the Hessian of the function \( f \) at a point \( \beta \) that belongs to an interior of \( S \), i.e. \( \beta \) is the local maximum of \( f \). This completes the proof of the lemma.

**Lemma 2** Let \( T: X \to X \) be a continuous map of a compact set \( S \) such that \( f(\beta) \geq f(\gamma) \) if \( T(\beta) \neq \beta \). Then all limits points of a sequence \( T^n(\beta) \) are fixed points of \( T \).

**Proof** Let \( a \) be a limit point of \( T^n(\beta), i = 1, 2, \ldots \) Then \( b = T(\beta) \) is a limit point for the sequence \( T^n(\beta) \). If \( b \neq a \) then we have
\[
f(T^n(b)) \leq f(T^n(\beta)) \leq f(T^{n+1}(\beta))
\] (14)

i.e. \( f(a) \leq f(\beta) \), This implies \( a = b \), Q.E.D.

**Lemma 3** Let \( f(\beta) \) be continuously differentiable on a compact subset \( X \subset \mathbb{R}^n \). Then for any \( q < 1 \) there exist a constant \( D_0 \) such that for any \( \beta \in X \) and \( D \geq D_0 \nabla^\beta \sum_{j=1}^{n} a_j \beta_j = 1 \)
\[
|T_D(\beta) - \beta| < q|\beta|
\] (15)

**Proof** We have
\[
T_D(\beta) = \frac{a_i \beta_i + D \beta_i}{\sum_{j=1}^{n} a_j \beta_j + D} = \frac{1/D a_i \beta_i + \beta_i}{\sum_{j=1}^{n} a_j \beta_j + D} =
\]
\[
\beta_i(1 + a_i/D) (1 - 1/D) \sum_{j=1}^{n} a_j \beta_j + O(1/D^2) =
\]
\[
\beta_i + \beta_i - \sum_{j=1}^{n} a_j \beta_j + O(1/D^2)
\] (16)

Therefore, there exists a sufficiently large \( D_0 \) such that
\[
|T_D(\beta) - \beta| < q|\beta|
\] (17)

for any \( D \geq D_0 \) where \( q < 1 \). Since \( f \) is continuously differentiable, there exist a small ball \( U_\beta \subset \mathbb{R}^n \) with the center at \( \beta \) such that for any \( \beta' \in U_\beta \)
\[
|T_D(\beta') - \beta'| < q|\beta'|
\] (18)

Since \( X \) is compact it can be covered by a finite number of such balls \( U_\beta \) and therefore one can choose \( D_0 \) and \( q \) satisfying the lemma conditions.

**Lemma 4** Let \( f(\beta) \) be continuously differentiable on a compact subset \( X \subset \mathbb{R}^n \). There exist a constant \( D_0 \) such that for any \( \beta \in X \) and \( D \geq D_0 \)
\[
f(T_D(\beta)) > f(\beta) \text{ if } T_D(\beta) \neq \beta
\] (19)

**Proof** If \( T_D(\beta) = \beta \) for some \( D \) then \( \beta \) is a critical point of Lagrangian \( \mathcal{L}(\beta, \lambda) = F - \lambda \sum \beta_j - 1 \) and for any \( D \) for which \( T_D(\beta) \) is defined \( T_D(\beta) = \beta \). Otherwise it was shown in [2] that there exists such large \( D \) that \( f(T_D(\beta)) > f(\beta) \).

Since \( f \) is continuously differentiable one can find a small open ball \( U_D \) of \( \beta \) such that \( f(T_{D'}(\beta')) > f(\beta') \) for any \( \beta' \in U_D \).

Since \( X \) is compact it can be covered by a finite set of open balls \( U_i \) such that for each ball \( U_i \) there is a \( D_i \) for which the lemma holds. Taking an upper limit of finite number of \( D_i \) proves the lemma.

**Theorem 1** Let \( f(\beta) \) be continuously differentiable on \( X \subset S \).

Let \( X = \bigcup C_i, i = 1, 2, \ldots \) where \( C_i \) are compact and \( C_i \subset C_{i+1} \). Assume also that for each \( C_i \) and any point \( \beta \in C_i \) there is an open ball \( U_i \subset \mathbb{R}^n \) at some center \( u_i \) such that \( \beta \in U_i \cap S \subset C_i \). Assume that for each \( \beta \in X \) there exists \( D_\beta \) such that for all \( D \geq D_\beta \) \( T_D(\beta) \in X \). Then there exists a transformation \( T: X \to X \) such that the following holds:

1. \( T(X) \subset X \)

2. For any \( \beta \in X \) such that \( T(\beta) \neq \beta \) we have:
\[
f(T(\beta)) > f(\beta)
\]

3. All limit points of a sequence \( \{T^n(\beta)\} \) are fixed points.

**Proof** Let define \( T \) inductively as follows. Let \( D_1 \) be such that for any \( D \geq D_1 \) and \( \beta \in C_1 \) we define \( T(\beta) = T(D_1) \). Assume that we defined \( T_k \) for \( C_k \). And let us define \( T \) on \( C_{k+1} \) as following. Let \( D_{k+1} \) be such that for any \( D \geq D_{k+1} \) and \( \beta \in C_{k+1} \) we define \( T_D(\beta) \) for \( \beta \in C_{k+1} \). Then we define \( T_{D_{k+1}}(\beta) = T_{D_1}(\beta) \) for \( \beta \in C_{k+1} \), and for \( \beta \in C_k \) we define \( T_{D_{k+1}}(\beta) = T_{D_{k+1}}(\beta) \). Since for any \( i < k \) \( T_{D_k}(\beta) = T_{D_{k+1}}(\beta) \) if \( \beta \in C_k \). All \( T \) give rise to the uniquely defined map \( T(\beta) = T_k(\beta) \) for \( \beta \in C_k \), \( k = 1, 2, \ldots \). The 3d statement of the theorem follows from Lemma 2. And the last statement of the theorem about local maximum follows from Lemma 1.

**Corollary 1** Let \( f \) be continuously differentiable everywhere in \( S \), except of finite number of points \( P \). Then the following holds for any \( \beta \in S_0 \setminus P \)

1. \( f(T^n(\beta)) > f(T^{n-1}(\beta)) \) if \( T^n(\beta) \neq T^{n-1}(\beta) \)

2. All limit points of \( T^n(\beta), n = 1, 2, \ldots \) are either critical points of the function \( f \) or belong to \( P \).

In addition, let a limit point \( \beta^* \) of \( T^n(\beta), n = 1, 2, \ldots \) belongs to \( S_0 \) and does not belong to \( P \). Let the Hessian of \( f \) at this point \( \beta^* \) is negative definite then this point \( \beta^* \) is the local maximum of \( f \).

**Proof** The corollary follows from the fact that the set \( X = S \setminus P \) can be represented as a union of compact sets \( C_i \) as described in Theorem 1. Namely, one can inductively build this set \( C_i \) as follows. For any point \( \beta_i \in X \) one can find an open ball \( U_i \) with a center at \( \beta_i \) such that this ball does not intersect \( P \). Let \( C_i \) be the closure of this ball. Next, we take any point \( \beta_i \in X \setminus P \) that lies outside of \( C_i \) and find another open ball \( U_{i+1} \) at the center of \( \beta_i \) that does not intersect \( C_i \) and \( P \). Then we define \( C_{i+1} \) as union of \( C_i \) and the closure of \( U_{i+1} \). Assume that we already constructed \( C_k \). Then we can for any point \( \beta_{k+1} \) find a ball \( U_{k+1} \) at the center of \( \beta_{k+1} \) such that \( U_{k+1} \) does not intersect \( C_k \) and \( P \). Then \( C_{k+1} \) is defined as a union of the closure of \( U_{k+1} \) and \( C_k \). Continue this process we build a set of \( C_k \) satisfying the conditions of Theorem 1 thereby proving the corollary.