A-FUNCTIONS: A GENERALIZATION OF EXTENDED BAUM-WELCH TRANSFORMATIONS TO CONVEX OPTIMIZATION

Dimitri Kanevsky, David Nahamoo, Tara N. Sainath, Bhuvana Ramabhadran, Peder A. Olsen

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

ABSTRACT

We introduce the Line Search A-Function (LSAF) technique that generalizes the Extended-Baum Welch technique in order to provide an effective optimization technique for a broader set of functions. We show how LSAF can be applied to functions of various probability density and distribution functions by demonstrating that these probability functions have an A-function. We also show that sparse representation problems (SR) that use l1 or combination of l1/l2 regularization norms can also be efficiently optimized through an A-function derived for their objective functions. We will demonstrate the efficiency of LSAF for SR problems through simulations by comparing it with Approximate Bayesian Compressive Sensing method that we recently applied to speech recognition.

Index Terms— Convex optimization, Extended Baum-Welch

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 [2] and [3], EBW technique was extended to estimating the parameter of Gaussian Mixture Models (GMMs) of HMMs under the MMI discriminative function for speech recognition problems. EBW's popularity, similar to Baum Welch algorithm, is based on its sinple recursion formula for updating the model parameters in the "Expectation-Maximization" (EM) fashion.

In [1] it is shown that the EBW recursion is a "growth" transformation, i.e. the value of the objective function increases at every iteration. The growth proof of GMMs was first given for arbitrary function of GMMs in [4] and in ([5],[6]) for rational functions. In [7], the EBW recursion formula for GMMs was recasted in a new form using the notion of Associated functions with an optimization process consisting of the following 3 steps: (1) find an associated function, (2) estimate the new parameters that maximize the associated function, and (3) do a search on the line connecting the current value of the parameters and the parameter values obtained in the step 2 to find the linear combination of parameters that yield the largest growth for the objective function.

In this paper we generalize the EBW technique beyond associated functions. We introduce the $\mathcal{A}$-function that will allow us to use the algorithm of [7] not only for the associated function of an objective function of GMM densities, but also associated functions of a variety of probability density and distribution functions such as Exponential, Poisson and Gamma functions. In addition we show that we can go beyond associated functions. We address the class of sparse representation problems where the objective function involves an l_1 norm term. We show how an $\mathcal{A}$-function can be used to optimize these l_1 norm based sparse representation problems. We call this extension of EBW, the Line Search A-functions (LSAF) optimization technique. We show that LSAF requires an $\mathcal{A}$-function that is strictly convex or concave. We show that LSAF has two steps: (1) optimization of the $\mathcal{A}$-function and (2) a line search on the line defined by current parameter values and the new parameter values obtained in Step 1. We show that the $\mathcal{A}$-function for probability distributions/densities can be constructed as an associated function as defined in [7]. We give examples of $\mathcal{A}$-functions for a few density/distribution functions. For optimization problems that involve sparse regularization metrics such as l_1 and squared l_1 norms we show a closed form representation of the $\mathcal{A}$-function.

We will also give a simple geometric proof that LSAF recursions result in a growth transformation. We demonstrate numerically the usefulness of LSAF approach vs. Approximate Baeysian Compressive Sensing method that was recently applied in an exemplar based sparse representation approach to speech recognition [8]. The rest of the paper is structured as follows. In Section 2 we introduce the notion of $\mathcal{A}$-function and describe LSAF. In Section 3 we show the use of $\mathcal{A}$-functions for general class of functions of density and distribution functions and in Section 4 we describe the LSAF process for sparse representation.

2. A-FUNCTION

2.1. Definition

Let $f(x) : \mathcal{U} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ be a real valued differentiable function in an open subset $\mathcal{U}$. Let $\mathbf{A}_f = \mathbf{A}_f(x, y) : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ be twice differentiable in $x \in \mathcal{U}$ for each $y \in \mathcal{U}$. We define $\mathbf{A}_f$ as an $\mathcal{A}$-function for f if the following properties hold.

1. $\mathbf{A}_f(x, y)$ is a strictly convex or strictly concave function of x for any $y \in \mathcal{U}$ (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 $z = g_y(x) = \mathbf{A}_f(x, y)$ and $z = f(x)$ at any $x = y \in \mathcal{U}$ are parallel to each other, i.e.

$$\nabla_x \mathbf{A}_f(x, y)|_{x=y} = \nabla_x f(x) \tag{1}$$

We will show that a general optimization technique can be constructed based on $\mathcal{A}$-function. We formulate 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 with the value that optimizes the $\mathcal{A}$-function, i.e. $\nabla_x \mathbb{A}_f(x, y)|_{x=\tilde{x}} = 0$. We call this technique Line Search A-Function (LSAF).

In this transformation process via an $\mathcal{A}$-function it is usually assumed that finding an optimum of an $\mathcal{A}$-function is "easier" than finding a (local) optimum of the original function f. Naturally, a desired outcome is for the equation $\nabla_x \mathbb{A}_f(x, y)|_{x=\tilde{x}} = 0$ to have

a closed form solution. We would like to draw the attention of the reader to two important observations:
1. The A-function and LSAF introduced in this paper and the technique suggested in [3] based on a weak-sense auxiliary function have key differences. First, the weak-sense auxiliary function is not required to be concave or convex. Therefore in general there is no guarantee of growth of the original function for update rules for parameters conducted via a weak-sense auxiliary function. Second, [3] does not suggest explicitly a line search as it is done in LSAF.
2. The LSAF optimization method via an A-f function is different from first order gradient optimization algorithms. This is because LSAF uses the optimum point of the A-function and therefore it adjusts the first order gradient direction of the desired objective function with additional term(s) based on second and higher order properties of the $\mathcal{A}$-function. We, therefore, conjecture that LSAF approximates higher order gradient optimization.

2.2. Growth Transformation Proof

We will now provide more details on why an $\mathcal{A}$-function gives a set of iterative update rules with a "growth" property (i.e. the value of the original function increases for the new parameters values). Let x_0 be some point in $\mathcal{U}$ and $\mathcal{U} \ni \tilde{x}_0 \neq x_0$ be a solution of $\nabla_x A(x, x_0)|_{x=\tilde{x}_0} = 0$ (it is the minimum of $\mathbf{A}_f(x, x_0)$ if $\mathbf{A}_f$ is concave and the maximum if $\mathbf{A}_f$ is convex). Let

$$x_1 = x(\alpha) = \alpha\tilde{x}_0 + (1-\alpha)x_0. \tag{2}$$

Then for sufficiently small $|\alpha| \neq 0$, $f(x(\alpha)) > f(x_0)$ where $\alpha > 0$ if $A(x, x_0)$ concave and $\alpha < 0$ if $A(x, x_0)$ convex. The formal proof of this is given in the Appendix A. In this section we demonstrate this growth property visually.
Let us define $\tilde{\mathbb{A}}_f(x, y) := \mathbb{A}_f(x, y) + f(y) - \mathbb{A}_f(y, y)$. We can easily see that $\tilde{\mathbb{A}}_f(y, y) = f(y)$. In addition since $f(y) - \mathbb{A}_f(y, y)$ does not depend on x, we can easily see that $\nabla_x \mathbb{A}_f(x, x_0)|_{x=\tilde{x}_0} = \nabla_x \tilde{\mathbb{A}}_f(x, x_0)|_{x=\tilde{x}_0} = 0$. As a result, $\tilde{\mathbb{A}}_f(x, y)$ is an $\mathcal{A}$-function of x that touches $f(x)$ at a point y. Having defined $\tilde{\mathbb{A}}_f(x, y)$, we consider several cases.
Case 1: auxiliary $\mathcal{A}$-function:
Let us assume that $f(x)$ is such that for any $y \in \mathcal{U}$ we can find $\tilde{\mathbb{A}}_f(x, y)$ as an $\mathcal{A}$-function of $f(x)$ such that $\tilde{\mathbb{A}}_f(x, y) \leq f(x)$ for all x. Such $\tilde{\mathbb{A}}_f$ is known as an auxiliary function. Further, if $f(x)$ is a likelihood function, $\tilde{\mathbb{A}}_f$ is the well known auxiliary function used in Expectation-Maximization algorithm. In this case one can set $\alpha = 1$ and if $\tilde{x}_0 \in \mathcal{U}$ then we have: $f(\tilde{x}_0) \geq \tilde{\mathbb{A}}_f(\tilde{x}_0, x_0) \geq \tilde{\mathbb{A}}_f(x_0, x_0) = f(x_0)$.

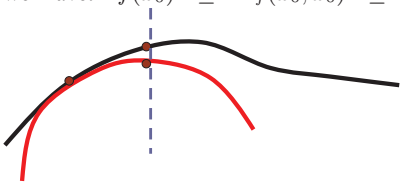

Fig. 1. Auxiliary case

In this figure the upper curve denotes the plot of the objective function $f : x \to \mathbb{R}$ and the curve in red, i.e. the convex lower curve, represents the $\mathcal{A}$-function $\tilde{\mathbb{A}}_f(., x_0) : x \to \mathbb{R}$. As it be can seen from this figure, for some $x_1 = \tilde{x}_0$ that maximizes $\tilde{\mathbb{A}}_f(x, x_0)$ we have $f(x_1) > f(x_0)$.

Case 2: concave $\mathcal{A}$-function:
Let us now assume that $\tilde{\mathbb{A}}_f(x, y)$ is convex in x. Further, let us assume that $x \in \mathbb{R}$, i.e. x is a one dimensional parameter. This is illustrated in Fig. 2.

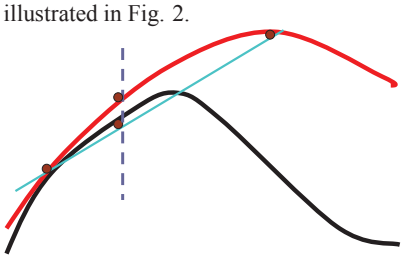

Fig. 2. Concave case

In this figure the lower curve denotes the plot of the objective function $f : x \to \mathbb{R}$ and the curve in red, i.e. the convex upper curve, represents the $\mathcal{A}$-function $\tilde{\mathbb{A}}_f(., x_0) : x \to \mathbb{R}$ (this curve can be below the objective function as well). As it can be seen from this figure, for some internal point x_1 between x_0 and $\tilde{x}_0$ we have $f(x_1) > f(x_0)$.

Case 3: convex $\mathcal{A}$-function: Here we assume that $\tilde{\mathbb{A}}_f(x, y)$ is concave in x. This case is illustrated in a Fig. 3.

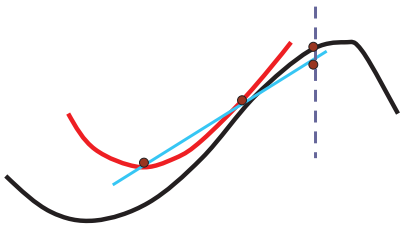

Fig. 3. Convex case

In this figure the lower curve denotes the plot of the objective function $f : x \to \mathbb{R}$ and the curve in red, i.e. the concave upper curve, represents the $\mathcal{A}$-function $\tilde{\mathbb{A}}_f(., x_0) : x \to \mathbb{R}$, (this curve can be below the objective function as well). As it be can seen from this figure, for some point x_1 outside of the segment $[x_0, \tilde{x}_0]$ we have $f(x_1) > f(x_0)$.

3. A-FUNCTIONS FOR GENERAL DISTRIBUTIONS/DENSITIES

3.1. Associated function

Let $\xi(x, \theta)$ be a density or distribution over a space $x \in \mathcal{X}$ with model parameters θ. Let $X_1^T = \{x_t \in \mathcal{X}\}, t = 1, ..T$ and consider a function $f(\{\xi_t(\theta)\})$ where $\xi_t = \xi_t(\theta) = \xi(x_t, \theta)$ for all $x_t \in X_1^T$. Let $c_t = c_t(\theta) = \xi_t \frac{\partial f(\{\xi_t\})}{\partial \xi_t}$ The following definition of an associated function was initially introduced for Gaussian densities in [7]. $Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\}) = \sum_t c_t(\theta_0) \log \xi_t(\theta)$. We will show that this associated function always satisfies the property stated in the equation (1). Indeed,

$$\nabla_\theta Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\})|_{\theta=\theta_0} = \sum c_t(\theta_0) \frac{\partial \log\{\xi_i(\theta)\}}{\partial \theta}|_{\theta=\theta_0}$$

$$= \sum \frac{c_t(\theta_0)}{\xi_i(\theta_0)} \frac{\partial \xi_t(\theta)}{\partial \theta}|_{\theta=\theta_0} = \sum \frac{\partial f(\{\xi_t(\theta_0)\})}{\partial \xi_t} \frac{\partial \xi_t}{\partial \theta}|_{\theta=\theta_0} =$$

$$\nabla_\theta f(\{\xi_t(\theta)\})|_{\theta=\theta_0}$$

3.2. Examples of A-functions

Assuming some conditions on c_t, we will now show that for the distributions described below, each $Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\})$ is strictly convex or concave. Hence, these specific associated functions are $\mathcal{A}$-functions.

$\mathcal{A}$-function for Exponential family: We define an exponential family as any family of densities on $\mathbb{R}^D$, parameterized by θ, that can be written $\xi(x,\theta) = \frac{\exp\{\theta^T\phi(x)\}}{Z(\theta)}$ where x is a D-dimensional base observation. The function $\phi : \mathbb{R}^D \to \mathbb{R}^d$ characterizes the exponential family. $Z(\theta) = \int_{\Xi} \exp\{\theta^T\phi(x)\}dx$ is the partition function, that provides the normalization necessary for $\xi(x,\theta)$. It was shown in [9] that $\log\xi(x,\theta)$ is convex and it is strictly convex if $Var[\phi(x)] \neq 0$. Therefore $Assoc_f$ is $\mathcal{A}$-function and defines the following growth model update parameters

$$\hat{\theta} = \alpha\tilde{\theta}_0 + (1-\alpha)\theta_0 \tag{3}$$

where $\tilde{\theta}_0$ is defined from the equality $\int_{x\in\mathcal{X}} \xi(\tilde{\theta}_0, x)\phi(x)dx = \frac{\sum c_t(\theta_0)\phi(x_t)}{\sum c_t(\theta_0)}$ and it is a solution of $\nabla_\theta \mathbb{A}_f(\theta,\theta_0)|_{\theta=\tilde{\theta}_0} = 0$. The proof of this is given in Appendix C. The recursion (3) is novel and different from ones that are used in [9].
$\mathcal{A}$-function for Poisson:

$$\xi(k,\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

$$\nabla^2 Assoc_f = \frac{-\sum c_t k_t}{\lambda^2}$$

has the same sign if $\lambda \neq 0$ and $\sum c_t k_t \neq 0$.

$\mathcal{A}$-function for Gamma

$$\xi(x,\alpha) = \frac{1}{\Gamma(\alpha)}x^{\alpha-1}e^{-x}$$

where $\alpha > 0$ and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ (see Appendix B for the proof).

$\mathcal{A}$-function for Gaussian: For $\theta = (\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R})$ we have

$$\xi(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_t-\mu)^2\}}{2\sigma^2}\right\}$$

Then

$$Assoc_f = Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\}) = -\sum c_t(\theta_0)\log(\sigma) -$$

$$\sum c_t(\theta_0)\frac{(x_t-\mu)^2}{2\sigma^2}$$

is strictly convex or concave in some neighborhood that contains μ_0, σ_0. Then we have the following updates for model parameters:

$$\hat{\mu} = \hat{\mu}(\alpha) = \alpha\tilde{\mu}_0 + (1-\alpha)\mu_0 \ , \hat{\sigma}^2 = \hat{\sigma}^2(\alpha) = \alpha\tilde{\sigma}_0^2 + (1-\alpha)\sigma_0^2 \tag{4}$$

where $\tilde{\mu}_0 = \frac{c_t(\theta_0)x_t}{\sum c_t(\theta_0)}$ and $\tilde{\sigma}_0^2 = \frac{c_t(x_t-\mu_0)^2}{\sum c_t}$. Recursions (4) were considered in ([10]) and called constrained line search. In [7] an EBW family of model updates was defined as those ones that coincide with $\hat{\mu} + o(\alpha^2), \hat{\sigma}^2 + o(\alpha^2)$. It was also shown there that EBW transformations

$$\bar{\mu} = \frac{\sum c_t(\theta_0)x_t + D\mu_0}{\sum c_t(\theta_0) + D} \tag{5}$$

$$\bar{\sigma}^2 = \frac{\sum c_t^2(\theta_0)x_t + D(\mu_0^2 + \sigma_0^2)}{\sum c_t(\theta_0) + D} - \bar{\mu}^2 \tag{6}$$

belongs to the EBW family for $\alpha = 1/D$, i.e. $\bar{\mu} = \hat{\mu} + o(\alpha^2)$ and $\bar{\sigma}^2 = \hat{\sigma}^2 + o(\alpha^2)$. Since $Assoc_f(\{\xi_t(\theta_0), \xi_t(\theta)\})$ is an $\mathcal{A}$-function, updates (4) are growth for sufficiently small $|\alpha|$.

4. SPARSE REPRESENTATIONS

Now we will look at the following sparse representation problem. Let $y \in \mathbb{R}^m$, $\beta \in \mathbb{R}^n$ and $H \in \mathbb{R}^{mn}$. Let us consider the constrained optimization problem

$$\min \parallel y - H\beta \parallel_R^2 \ \text{ s.t. } \ \parallel \beta \parallel_1 < \epsilon \tag{7}$$

In many practical application it is useful to add an l_2 regularization term and therefore we consider the problem:

$$\min \parallel y - H\beta \parallel_R^2 + \parallel \beta - \beta_0 \parallel_{P_0}^2 \ \text{ s.t. } \ \parallel \beta \parallel_1 < \epsilon$$

Using $\parallel y - H\beta \parallel_R^2 + \parallel \beta - \beta_0 \parallel_{P_0}^2 = \parallel \beta - \beta_1 \parallel_{P_1}^2$ we can represent this problem as $\min \parallel \beta - \beta_1 \parallel_{P_1}^2$ s.t. $\parallel \beta \parallel_1 < \epsilon$ where P_1 is assumed to be positive-definite. We can now represent the equation (7) by solving the problem

$$\text{Minimize } F(\beta) = \parallel \beta - \beta_1 \parallel_{P_1}^2 + \parallel \beta \parallel_1^i / \sigma^2 \tag{8}$$

and define the $\mathcal{A}$-function as:

$$\mathbb{A}(\beta, \beta^*) = \parallel \beta - \beta^* \parallel_{P_1}^2 + \{sign(\beta^*)\beta\}^i / \sigma^2 \tag{9}$$

where $i = 1$ (laplacian) or $i = 2$ (squared l_1 norm). In Appendix D we show that $\mathbb{A}(\beta,\beta^*)$ is $\mathcal{A}$-function of $F(\beta)$. According to the definition of the $\mathcal{A}$-function, we consider $\mathbb{A}(\beta,\beta^*)$ and $F(\beta)$ in an open domain where they are both differentiable and construct an update of parameters when the extremum of $\mathbb{A}(\beta,\beta^*)$ belongs to this domain. Our open domain excludes the origin, i.e. $\beta = 0$, and all coordinates of $\beta(j) = 0$. If some coordinates of β approach 0 we can remove them by reducing the dimension of the problem. Using LSAF we have the recursion formula:

$$\beta_k = \alpha\tilde{\beta}_{k-1} + (1-\alpha)\beta_{k-1}$$

Previously we introduced an ABCS algorithm [8] for compressive sensing with squared $l_1$norm, i.e. $i = 2$, of the equation (8). Analysis of various regularization penalties for speech classification problems was given in [11]. The ABCS method gives a solution of (8) via the recursion: $\tilde{\beta}_{k-1} = \arg\max_\beta A(\beta, \beta_{k-1})$. We used this $\tilde{\beta}_{k-1}$ recursion solution of 8) in many experiments succesfully, i.e. with good sparse signal recovering. We did not provide a growth proof. Numerical experiments show that for a suitable choice of α β_k converges faster to a solution of (8) than the one obtained through the ABCS recursion. One can expect that LSAF with appropriate choices of α is more efficient that ABCS (that corresponds to $\alpha = 1$).

Simulation:
We run numerical experiments in which we recovered a parameter vector β in (7). In our example the signal $\beta \in \mathbb{R}^{256}$ is assumed to be a sparse parametric vector. The signal support consists of 10 elements $\beta(i) \neq 0$. In our simulation the non-zero $\beta(i)$ values and the index are uniformly sampled over $\beta(i)$ $U[-10,10]$ and i $U_i[1,256]$, respectively. The sensing matrix $H \in \mathbb{R}^{72x256}$ consists of entries that are sampled according to $\mathcal{N}(0, 1/72)$. In Figure 4 illustrates the actual signal (spikes), the recovered signal via LSAF (squares) and the recivered signal via ABCS (+) after 100 iterations. As one can see LSAF is more accurately. In order for ABCS to get the same level of accuracy that LSAF has one needs to run 1000 iterations of ABCS.

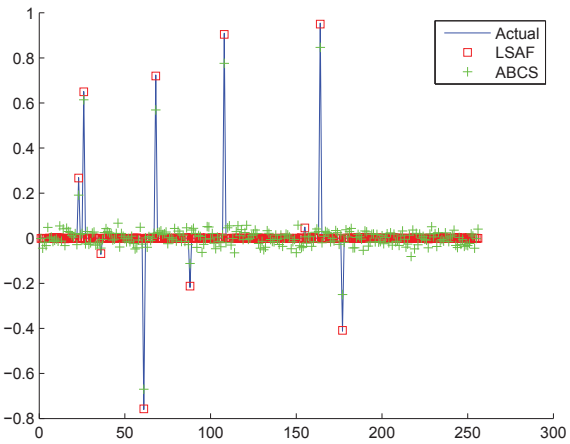

Fig. 4. Simulation: ABCS vs. LSAF, 100 iterations

5. CONCLUSIONS

In this paper, we introduced LSAF as a general optimization method that extends EBW transformations to a broader set of objective functions. We introduced the novel concept of $\mathcal{A}$-functions. Using $\mathcal{A}$-functions, we showed that LSAF algorithm can efficiently update the parameters of a function to increase its value. We demonstrated that $\mathcal{A}$-functions can be constructed as associated functions for various families of densities and distributions such as - Exponential models, Poisson, Gamma functions. Finally, we showed the utility of LSAF for various sparse representation objective functions. We showed that LSAF is more efficient than ABCS for sparse representation. We plan to investigate the rate of convergence of LSAF for various applications.

6. APPENDICES

***A**: Growth proof of parameter updates via $\mathcal{A}$-function*

Here we prove that update in (2) is growth for the function $f(x)$. Let $\mathbb{R}^{n+1}$ be a coordinate space that contains the graph $\{\mathbb{R}^n, f(\mathbb{R}^n)\}$. Let l_{x_0} be a tangent line to $f(x)$ at x_0 that passes through a plane that contains the line l_{x_0} and a line through $(\tilde{x}_0, \mathcal{A}_f(\tilde{x}_0, x_0)) \in \mathbb{R}^{n+1}$ and $(\tilde{x}_0, 0) \in \mathbb{R}^{n+1}$. In what follows we can restrict $\mathcal{A}_f$ and f to this plane. l_{x_0} also tangents to $g(x) = \mathcal{A}_f(x, x_0)$ at x_0. Let us consider the case when $\mathcal{A}_f(x, x_0)$ is a concave function in x. Then for sufficiently small ϵ we have $g(x_0 + \epsilon l_{x_0}) > g(x_0)$. Since $g(x)$ and $f(x)$ have the same tangent line l_{x_0} at x_0 for sufficiently small ϵ $f(x_0 + \epsilon l_{x_0}) > f(x_0)$. (This follows from the fact that l_{x_0} is not parallel to a coordinate hyperplane $\mathbb{R}^n$ since x_0 is not an extremum of $g(x)$)). Let $\alpha > 0$ be such that $x(\alpha) = (1-\alpha)\tilde{x}_0 + \alpha x_0 = x_0 + \epsilon l_{x_0}$. Then $f(x(\alpha)) > f(x_0)$. Similarly one can prove a statement for $g(x)$ convex.

Remark: We conjecture that one can relax the condition of concavity/convexity in the definition of the $\mathcal{A}$-function and require it to be unimodal in x for any $y \in \mathcal{U}$. This allows to extend the basic theory of this paper to other distribution/density functions for which the $\mathcal{A}$-function is not strictly concave/convex.

***B**: Proof that an associated function for Gamma function is $\mathcal{A}$-function*

$\xi(x, \alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x}$ where $\alpha > 0$ and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$

$\mathcal{A}$-function for f can be computed as $\mathbb{A}_f = Assoc_f(\{\xi_t(\alpha_0), \xi_t(\alpha)\}) = \sum c_t(\alpha_0) log\{\xi_i(\alpha)) = \sum_t c_t\{\log x^{\alpha-1} e^{-x} - \log\Gamma(\alpha)\}$ In order to prove convexity/concavity property one need to compute a second derivative of $\mathbb{A}_f$ and see if its sign preserves. We have: $\mathbb{A}'_f = \sum_t c_t \left\{\log x - \frac{\Gamma(\alpha)'}{\Gamma(\alpha)}\right\} = const - \frac{1}{\alpha} - \sum_t c_t \sum_{n=1}^\infty \frac{\alpha}{n(\alpha+n)}$ Taking another derivative gives: $\mathbb{A}''_f = -\sum_t c_t \sum_{n=1}^\infty \frac{\alpha}{n(\alpha+n)} = \sum_t c_t \sum_{n=1}^\infty \frac{1}{(\alpha+n)^2} > 0$ if $\sum_t c_t \neq 0$

***C**: Calculation of $\nabla_\theta \mathbb{A}_f(\theta, \theta_0)|_{\theta=\tilde{\theta}_0} = 0$ for exponential functions.*

$\nabla \log \xi_t(\theta) = \frac{\nabla \xi_t(\theta)}{\xi_t(\theta)}$ and $\nabla \xi_t(\theta) = \xi_t(\theta)\nabla(\theta^T \phi(x_t)) - \xi_t(\theta)\frac{\nabla Z(\theta)}{Z(\theta)} = \xi_t(\theta)\left[\phi(x_t) - \frac{\int_{x\in\mathcal{X}} \nabla(\exp\{\theta^T\phi(x)\})dx}{Z(\theta)}\right]$ Therefore $\nabla \log \xi_t(\theta) = \phi(x_t) - \frac{\int_{x\in\mathcal{X}} \exp\{\theta^T\phi(x)\}\phi(x)dx}{Z(\theta)} = \phi(x_t) - \int_{x\in\mathcal{X}} \xi(\theta, x)\phi(x)dx$ - set it equal to zero and substitute into definition of $Assoc_f$ function to get the equality below (3).

***D**: Proof that (9) is an $\mathcal{A}$-function for (8).*

First, we have to prove the property stated in the equation (1). We have:

$$\nabla_\beta F(\beta)|_{\beta=\beta_1} = \nabla_\beta \left(\| \beta - \beta_1 \|^2_{P_1}\right)|_{\beta=\beta_1} +$$

$$\frac{\nabla_\beta \|\beta\|_1^i|_{\beta=\beta_1}}{\sigma^2} = \nabla_\beta \mathbb{A}(\beta, \beta^*)|_{\beta=\beta_1}$$

since $\nabla_\beta \| \beta \|_1^i |_{\beta=\beta_1} = i \| \beta_1 \|_1^{i-1} sign(\beta_1) = \nabla_\beta\{sign(\beta_1)\beta\}^i|_{\beta=\beta_1}$ (note that $\beta = 0$ is not in the open domain U and $\nabla sign(\beta)$ in open domain U is zero everywhere).

Next, we have $\nabla^2_\beta \mathbb{A}(\beta, \beta_1)|_{\beta=\beta_1}$ is equal $2P_1$ for $i = 1$ or $2P_1$ + a semi-definite positive matrix for $i = 2$ and in both cases the Hessian matrix is positive definite. Therefore (9) is strictly convex. Q.E.D.

7. REFERENCES

[1] 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.

[2] Y. Normandin, "An improved MMIE training algorithm for speaker-independent, small vocabulary, continuous speech recognition," in *Proc. ICASSP*, 1991.

[3] D. Povey, *Discriminative Training for Large Vocabulary Speech Recognition*, Ph.D. thesis, University of Cambridge, 2003.

[4] D. Kanevsky, "Extended Baum Transformations For General Functions, II," Tech. Rep. RC23645(W0506-120), Human Language Technologies, IBM, 2005.

[5] T. Jebara, "On reversing Jensen's inequality," in *Proc. NIPS*, 2002.

[6] S. Axelrod, V. Goel, P. Gopinath R., Olsen, and K. Visweswariah, "Discriminative estimation of subspace constrained gaussian mixture models for speech recognition," *IEEE Transactions in Speech and Audio Processing*, vol. 15, no. 1, pp. 172 – 189, 2007.

[7] D. Kanevsky, T. Sainath, B. Ramabhadran, and D. Nahamoo, "Generalization of Extended Baum-Welch Parameter Estimation for Discriminative Training and Decoding," in *Proc. Interspeech*, 2008.

[8] T. N. Sainath, A. Carmi, D. Kanevsky, and B. Ramabhadran, "Bayesian Compressive Sensing for Phonetic Classification," in *In Proc. ICASSP*, 2010.

[9] V. Goel and P. Olsen, "Acoustic modeling using exponential families," October 2009, Proc. Interspeech.

[10] C. Liu, P. Liu, H. Jiang, Y. Liu, F. Soong, and R. Wang, "A Constrained Line Search Optimization for Discriminative Training in Speech Recognition," in *Proc. ICASSP*, 2007.

[11] D. Kanevsky, T. Sainath, B. Ramabhadran, and D. Nahamoo, "An Analysis of Sparseness and Regularization in Exemplar-Based Methods for Speech Classification," in *Proc. Interspeech*, 2010.