

Convergence Theorems for Generalized Alternating Minimization Procedures

Asela Gunawardana

*Microsoft Research
One Microsoft Way
Redmond, WA 98052, U.S.A.*

ASELAG@MICROSOFT.COM

William Byrne

*University of Cambridge
Department of Engineering
Trumpington Street
Cambridge, CB2 1PZ, U.K.*

WJB31@CAM.AC.UK

Editor: Michael I. Jordan

Abstract

The EM algorithm is widely used to develop iterative parameter estimation procedures for statistical models. In cases where these procedures strictly follow the EM formulation, the convergence properties of the estimation procedures are well understood. In some instances there are practical reasons to develop procedures that do not strictly fall within the EM framework. We study EM variants in which the E-step is not performed exactly, either to obtain improved rates of convergence, or due to approximations needed to compute statistics under a model family over which E-steps cannot be realized. Since these variants are not EM procedures, the standard (G)EM convergence results do not apply to them. We present an information geometric framework for describing such algorithms and analyzing their convergence properties. We apply this framework to analyze the convergence properties of incremental EM and variational EM. For incremental EM, we discuss conditions under these algorithms converge in likelihood. For variational EM, we show how the E-step approximation prevents convergence to local maxima in likelihood.

Keywords: EM, variational EM, incremental EM, convergence, information geometry

1. Introduction

The expectation-maximization (EM) algorithm (Dempster et al., 1977) for maximum likelihood estimation (Fisher, 1922; Wald, 1949; Lehmann, 1980) is one of the most widely used parameter estimation procedures in statistical modeling. It is clear why the algorithm is attractive to researchers building statistical models. The algorithm has an elegant formulation and when it is applied to appropriate model architectures it yields parameter update procedures that are easy to derive and straightforward to implement. These parameter estimates yield increasing likelihood over the training data, and the convergence behavior of this process is well understood.

EM also has acknowledged shortcomings. It can be slow to converge or even intractable for some combinations of models and training data sets, and there are also model architectures for which the straightforward application of EM yields update procedures that do not have closed form expressions. As a result, many improvements and extensions of EM have been developed (e.g., Meilijson, 1989; Salakhutdinov et al., 2003). Incremental EM (Neal and Hinton, 1998) and vari-

ational EM (Jordan et al., 1999) are specific examples we will address in the sequel. Such extensions improve various aspects of EM, such as rate of convergence and computational tractability. However, classical (generalized) EM convergence analyses such as those of Wu (1983) and Boyles (1983) do not apply to many of these variants, and in many cases their convergence behavior is poorly understood.

We propose the generalized alternating minimization (GAM) framework with the goal of understanding the convergence properties of a broad class of such EM variants. It is based on the interpretation of EM as an alternating minimization procedure as described by Csiszár and Tusnády (1984) and later by Byrne (1992) and Amari (1995). We will show that this alternating minimization procedure can be extended in a manner analogous to the manner in which generalized EM (GEM) extends the M step of EM. We then apply a convergence argument similar to that of Wu (1983) to GAM algorithms, characterizing their convergence. This will show that GAM algorithms are a further generalization of GEM algorithms which are no longer guaranteed to increase likelihood at every iteration, but nevertheless retain convergence to stationary points in likelihood under fairly general conditions.

In practice, an iteration of EM consists of an E step which calculates sufficient statistics under the posterior distribution of the most recent model estimate, followed by an M step which generates a new model estimate from those statistics. In contrast, many variants redefine the E step to use sufficient statistics calculated under other distributions. For example, an approximation to this posterior distribution is used in variational EM (Jordan et al., 1999), and statistics from the posterior distributions of previous estimates are carried over in incremental EM (Neal and Hinton, 1998). Existing (G)EM convergence results do not apply because the E step in such variants is modified to use other “generating distributions” for computing the sufficient statistics. In order to describe such variants where the generating distribution is not necessarily the posterior distribution under the current model, GAM keeps track of both the current model and the distribution generating the statistics used for computing the next model estimate. While EM algorithms generate sequences of parameters, GAM algorithms generate sequences of parameters paired with these generating distributions.

We use the GAM framework to analyze the convergence behavior of incremental EM (Neal and Hinton, 1998) and variational EM (Jordan et al., 1999). We show that incremental EM converges to stationary points in likelihood under mild assumptions on the model family. The convergence behavior of variational EM is more complex. We do show how GAM convergence arguments can be used to guarantee the convergence of a broad class of variational EM estimation procedures. However, unlike incremental EM, this does not guarantee convergence to stationary points in likelihood. On the contrary, we show that fixed points under variational EM cannot be stationary points in likelihood, except in the degenerate case when the model family is forced to satisfy the constraints that define the variational approximation itself.

In Section 2, we review how the EM algorithm results from alternating minimization of the information divergence. First, the divergence from the current model to a family of distributions of a certain form is minimized to give a generating distribution. Then, the divergence from this distribution to the model family is minimized to give the next model. We then show that extensions of the E step such as those mentioned above involve choosing “generating distributions” that do not minimize the divergence. In GAM, the E and the M steps need only reduce—and not minimize—the divergence. In fact, the steps need not reduce the divergence individually, but may do so when applied in succession. As in EM, the modeling assumptions are represented in the parameterization of

the models. Additionally, GAM explicitly represents the approximations used in estimation by imposing constraints on the generating distributions. In pursuing this formulation we were influenced by the work of Neal and Hinton (1998) which uses generating distributions to introduce several EM variants. Our intention is to extend their analysis and provide convergence results for the algorithms they and others propose.

Understanding the convergence behavior of these variants requires the analysis of joint sequences of both parameters and their corresponding generating distributions. In Section 3 we present such an analysis. Our main convergence theorem gives conditions under which GAM procedures converge to EM fixed points. We draw on the previous work of Wu (1983) which uses results from nonlinear programming to give conditions under which (G)EM procedures converge to stationary points in likelihood, as well as the work of Csiszár and Tusnády (1984) which gives an information geometric treatment of (G)EM procedures as generating joint sequences of generating distributions and parameters. Csiszár and Tusnády (1984) also provide a convergence analysis that complements the original results of Wu (1983). However neither of the approaches generalize to EM extensions that extend the E step.

In Section 4 we apply our convergence results to incremental EM and show that although the algorithm is non-monotonic in likelihood, it does converge to EM fixed points under very general conditions. Note that Neal and Hinton (1998) have already shown that incremental EM gives non-increasing divergence (non-decreasing free energy) and that local minima in divergence (local maxima in free energy) are local maxima in likelihood. However, as we show in Section 3, this is insufficient to conclude that incremental EM converges to local maxima in likelihood, and the further analysis that is necessary is presented here. In Section 5 we apply a similar analysis to variational EM to show that convergence to EM fixed points occurs only in degenerate cases. We then conclude with some discussion in Section 6.

2. EM and Generalized Alternating Minimization

We adopt the view of the EM algorithm as an alternating minimization procedure under the information geometric framework as developed by Csiszár and Tusnády (Csiszár and Tusnády, 1984; Csiszár, 1990). This framework allows an intuitive understanding of the algorithm, and is easily extended to cover many EM variants of interest. In Section 2.1, we briefly review the EM algorithm as derived within this framework to set the groundwork for the convergence analysis of later sections. In particular we show how EM can be derived as the alternating minimization of the information divergence between the model family and a set of probability distributions constrained to be concentrated on the training data. In Section 2.2, we then extend this alternating minimization framework to generalized alternating minimization (GAM) algorithms, which are EM variants that allow extensions of the E step, in addition to the M step extensions allowed by GEM algorithms. We conclude our introduction to GAM algorithms by discussing how the GAM framework is applied to algorithms of interest in Section 2.3.

2.1 EM as Alternating Minimization

The EM algorithm, when viewed as an alternating minimization procedure, minimizes a Kullback-Leibler type divergence between a *model family* (or equivalently a parameter family) and a *desired family* of probability distributions (these are the previously mentioned generating distributions).

Let the pair of random variables X and Y be related through a function mapping X to Y . That is, X is the complete random variable and Y is the incomplete, or observed, random variable (Dempster et al., 1977; McLachlan and Krishnan, 1997). Often, X is composed of an observed and a hidden part, and Y is composed of only the observed part. We adopt the “complete”/“incomplete” variable terminology of Dempster et al. (1977) rather than the “observed”/“hidden” variable terminology that is also commonly used. The model family \mathcal{P} is defined as the set of parameterized models $P_{X;\theta}$ obtained when θ ranges over the parameter family Θ . For simplicity, we make the following assumptions

- (Q1) The complete variable X is discrete-valued.
- (Q2) $p_X(x; \theta) > 0$ for all $\theta \in \Theta$ and for all values x taken on by X . That is, the support of the models does not depend on the parameter.
- (Q3) The p.d.f. $p_X(x; \theta)$ is continuous in θ .

These technical restrictions can be relaxed to allow continuous variables (Gunawardana, 2001). The difficulty faced in doing so is that continuous models assign zero probability to the training samples; Csiszár and Tusnády (1984) show how this problem can be circumvented by the introduction of an appropriate family of dominating measures.

The desired family \mathcal{D} is defined as the set of all probability distributions Q_X that assign probability one to the observation \hat{y} of Y :

$$\mathcal{D} \triangleq \{Q_X : q_Y(\hat{y}) = 1\}$$

where Q_Y is obtained by marginalizing Q_X . Thus, desired distributions $Q_X \in \mathcal{D}$ have the property that $Q_X = Q_{X|Y=\hat{y}}$. These probability distributions are “desired” in the sense that they exemplify the maximum likelihood estimation criterion by assigning the highest possible probability to the observed data \hat{y} . Note that multiple training examples are treated by considering the sequences $X = (X^{(1)}, \dots, X^{(n)})$ and $Y = (Y^{(1)}, \dots, Y^{(n)})$ together with suitable i.i.d. assumptions.

Since we will be concerned with estimating parameterized models $P_{X;\theta}$, we define the Kullback-Leibler information divergence (Liese and Vajda, 1987) between a desired distribution $Q_X \in \mathcal{D}$ and a parameter $\theta \in \Theta$ through

$$D(Q_X || P_{X;\theta}) = \sum_x q_X(x) \log \frac{q_X(x)}{p_X(x; \theta)}. \tag{1}$$

Note that the divergence is finite for all desired distributions $Q_X \in \mathcal{D}$ and all parameters $\theta \in \Theta$ because of our simplifying assumption about the support of models $P_{X;\theta}$. This implies that the divergence is continuous over all $(Q_X, \theta) \in \mathcal{D} \times \Theta$.

Csiszár and Tusnády (1984) show that the EM algorithm can be derived as alternating minimization under the information divergence, as follows (see Figure 1):

Forward Step: Find the desired distribution $Q_X^{(t+1)}$ that minimizes the divergence from the previous parameter $\theta^{(t)}$:

$$Q_X^{(t+1)} = \operatorname{argmin}_{Q_X \in \mathcal{D}} D(Q_X || P_{X;\theta^{(t)}}).$$

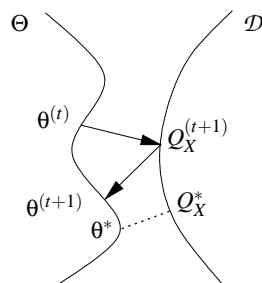


Figure 1: A schematic representation of an iteration of the alternating minimization procedure. The square of the distance between a point in Θ and a point Q_X in \mathcal{D} indicates the divergence between them. The arrows indicate projection under this divergence.

Backward Step: Find the parameter $\theta^{(t+1)}$ that minimizes the divergence to $Q_X^{(t+1)}$:

$$\theta^{(t+1)} \in \operatorname{argmin}_{\theta \in \Theta} D(Q_X^{(t+1)} || P_{X;\theta}). \quad (2)$$

In the language of Csiszár (1975), $Q_X^{(t+1)}$ is the *I-projection* of $P_{X;\theta^{(t)}}$ onto \mathcal{D} and is uniquely found as $Q_X^{(t+1)} = P_{X|Y=\hat{y};\theta^{(t)}}$. The EM algorithm (Dempster et al., 1977; Wu, 1983) can be recovered easily by substituting the I-projection into equation (2) and expanding the divergence using equation (1), to obtain

$$\theta^{(t+1)} \in \operatorname{argmax}_{\theta \in \Theta} \mathbf{E}_{P_{X|Y}} \left[\log p_X(X; \theta) \mid \hat{y}; \theta^{(t)} \right]. \quad (3)$$

Note that we use the notation $\theta^{(t+1)} \in \operatorname{argmin} D(Q_X^{(t+1)} || P_{X;\theta})$ instead of $\theta^{(t+1)} = \operatorname{argmin} D(Q_X^{(t+1)} || P_{X;\theta})$ because the backward step may not be unique.

We distinguish between the forward and backward steps of the alternating minimization procedure and the E and M steps of the EM procedure, as they are subtly different. The E step corresponds to computing the (conditional) expected log likelihood (EM auxiliary function) under the result of the forward projection. In practical implementations, the auxiliary function is not computed explicitly in the E step – the expected sufficient statistics are all that need be computed. Thus, the E step corresponds to taking an expectation under the distribution found in the forward step. The backward projection minimizes the divergence from the result of the forward projection, while the M step maximizes the expected log likelihood computed in the E step (or alternatively, finds parameters such that the sufficient statistics of the resulting model match those computed in the E step).

2.2 Generalized Alternating Minimization

There are many effective learning algorithms originally motivated by EM but which cannot be described using the formulation described above, or equivalently, using the original formulation of Dempster et al. (1977), because they generalize either the forward or the backward step. Two examples of such procedures are incremental EM (Neal and Hinton, 1998) and variational EM (Jordan

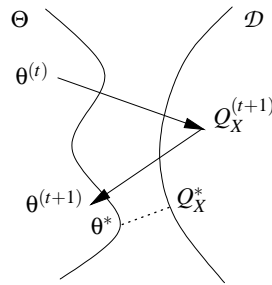


Figure 2: A schematic representation of an E step extension allowed by GAM algorithms corresponding to the M step extension of the GEM algorithm. In contrast to Figure 1, both the E and M steps reduce the divergence rather than minimizing it.

et al., 1999). We are interested in extending the alternating minimization formulation to such variants by relaxing the requirement that the forward and backward steps perform exact minimization over the families of distributions. These generalized estimation steps are described as follows.

Generalized Forward Step: Find any desired distribution $Q_X^{(t+1)}$ that reduces divergence from the previous parameter $\theta^{(t)}$:

$$Q_X^{(t+1)} : D(Q_X^{(t+1)} || P_{X;\theta^{(t)}}) \leq D(Q_X^{(t)} || P_{X;\theta^{(t)}}).$$

Generalized Backward Step: Find a parameter $\theta^{(t+1)}$ that reduces the divergence to $Q_X^{(t+1)}$:

$$\theta^{(t+1)} : D(Q_X^{(t+1)} || P_{X;\theta^{(t+1)}}) \leq D(Q_X^{(t+1)} || P_{X;\theta^{(t)}}). \tag{4}$$

Generalizations of the backward step correspond to the well known GEM algorithms. We allow similar generalization of the forward step. We refer to algorithms that consist of alternating application of such generalized forward and backward steps as generalized alternating minimization (GAM) algorithms. Thus, GAM algorithms allow for the expectation in the EM auxiliary function (equation (3)) to be found under the distribution $Q_X^{(t+1)}$ rather than $P_{X|Y=y; \theta^{(t)}}$. $Q_X^{(t+1)}$ is not chosen arbitrarily; it must be closer to $P_{X; \theta^{(t)}}$ than the desired distribution $Q_X^{(t)}$ used at the previous iteration. The effect of GAM iterations is to generate sequences of paired distributions and parameters $(Q_X^{(t)}, \theta^{(t)})$ that satisfy

$$D(Q_X^{(t+1)} || P_{X;\theta^{(t+1)}}) \leq D(Q_X^{(t)} || P_{X;\theta^{(t)}}).$$

Thus, we examine generalizations that are composed of forward and backward steps that reduce the divergence, as shown schematically in Figure 2.

2.3 Why GAM

As shown by Jordan et al. (1999), the variational EM algorithm is best described as an alternating minimization between a set of parameterized models and a set of variational approximations to the

posterior. This corresponds to extending the forward step to be a projection onto a subset of \mathcal{D} which satisfies additional constraints (namely, belonging to a given parametric family), rather than a projection onto \mathcal{D} itself.

In the following example, which follows Jordan et al. (1999), we describe how the mean field approximation to the E step arises by further constraining the desired family \mathcal{D} .

Example 1 *In the case of a Boltzmann machine, we have binary r.v.s $X = S = (S_1, \dots, S_n)$ modeled by the parametric family*

$$p_S(s; \theta) = \frac{e^{\sum_{i < j} \theta_{ij} s_i s_j + \sum_i \theta_{i0} s_i}}{z(\theta)}$$

where $z(\theta)$ ensures $P_{S; \theta}$ is properly normalized. Suppose the nodes $1, \dots, n$ of the Boltzmann machine are partitioned into a set of evidence nodes E and a set of hidden nodes H , so that $Y = S_E \triangleq (S_i)_{i \in E}$. Then, given observations \hat{s}_E of the evidence nodes, the forward step for EM estimation of the Boltzmann machine is as follows.

Forward Step: *Finding the desired distribution*

$$Q_X^{(t+1)} = \underset{Q_X \in \mathcal{D}}{\operatorname{argmin}} D(Q_X || P_{X; \theta^{(t)}})$$

gives

$$q_{S_H, S_E}^{(t+1)}(s_H, s_E) = 1_{\hat{s}_E}(s_E) p_{S_H | S_E}(s_H | \hat{s}_E; \theta^{(t)})$$

where $1_{\hat{s}_E}(s_E) = 1$ when $s_E = \hat{s}_E$ and 0 otherwise.

Note that a closed-form solution for the backward step is not generally available, but convergent algorithms can be obtained using gradient descent or iterative proportional fitting (Darroch and Ratcliff, 1972; Byrne, 1992). While the forward step can in principle be carried out exactly, this computation quickly becomes intractable as the number of states increases. In particular, direct computation of $p_{S_H | S_E}(s_H | \hat{s}_E; \theta^{(t)})$ using Bayes rule involves a summation over all possible values of the hidden nodes s_H .

To get around this we define a subset of \mathcal{D} consisting of mean field approximations to $Q_{S_H | S_E}$. That is, we define \mathcal{D}_{MF} to be those distributions in \mathcal{D} whose p.d.f. has the parametric form

$$q_S(s; \mu) = 1_{\hat{s}_E}(s_E) \prod_{i \in H} \underbrace{\mu_i^{s_i} (1 - \mu_i)^{1 - s_i}}_{q_{S_i; \mu_i}}$$

where each μ_i takes values in $[0, 1]$. Thus the members of \mathcal{D}_{MF} allow no dependencies between nodes. It follows that a distribution $Q_S \in \mathcal{D}_{MF}$ is fully specified by its parameter μ and the training observations \hat{s}_E .

The forward step can then be replaced by an approximate forward step, which is now a minimization over the variational parameter μ for fixed $\theta^{(t)}$:

Approximate Forward Step: *An (approximate) desired distribution*

$$Q_X^{(t+1)} \in \underset{Q_X \in \mathcal{D}_{MF}}{\operatorname{argmin}} D(Q_X || P_{X; \theta^{(t)}})$$

with *p.d.f.*

$$q_S^{(t+1)}(s) = 1_{s_E}(s_E) \prod_{i \in H} q_{S_i}(s_i; \mu_i^{(t+1)})$$

is chosen by finding a variational parameter

$$\mu^{(t+1)} \in \underset{\mu}{\operatorname{argmin}} D(Q_{S;\mu} || P_{S;\theta^{(t)}}).$$

As described by Jordan *et al.* (1999), this can be done directly, without needing to compute $P_{S_H|S_E;\theta^{(t)}}$, by solving the nonlinear system of mean field equations

$$\mu_i^{(t+1)} = \sigma \left(\sum_j \theta_{ij}^{(t)} \mu_j^{(t+1)} + \theta_{i0} \right),$$

where $\sigma(\cdot)$ is the logistic function. Note that this simplification results from the careful crafting of the parametric form imposed on \mathcal{D}_{MF} .

It can be seen that this variational EM variant is easily described in terms of minimizing the divergence between a constrained family of desired distributions and a model family. The approximate forward step in this example is a generalization of the usual I-projection onto \mathcal{D} , and the resulting algorithm is therefore a GAM procedure.

3. GAM Convergence

In this section, we describe our main result – a theorem which characterizes the convergence of GAM procedures. As the preceding example shows, some EM variants of interest are GAM procedures but not GEM procedures. This means that their convergence behavior may be different from what the familiar convergence properties of (G)EM would suggest. In particular, monotonic increase in likelihood and convergence to local maxima (technically, stationary points) in likelihood may no longer hold. This may happen even when the divergence is non-increasing, and when stationary points of the likelihood are fixed points of the GAM procedure. We begin with a simple toy example where this can easily be seen.

Example 2 Let the complete random variable $X = (X_1, X_2)$ represent the result of tossing a coin twice. That is, X_1, X_2 are i.i.d., with X_i taking the value 1 with probability θ and 0 with probability $1 - \theta$. Let the incomplete random variable Y encode whether the result seemed “fair” or not. It takes on the value 1 if X takes on the values (0,1) or (1,0), and takes on the value 0 otherwise. Suppose the observation \hat{y} of Y is $\hat{y} = 0$. In this simple case, the complete data likelihood is given by

$$p_X(x; \theta) = \theta^{x_1+x_2} (1 - \theta)^{2-(x_1+x_2)}$$

and the incomplete data likelihood is given by

$$\begin{aligned} p_Y(\hat{y}; \theta) &= p_Y(0; \theta) \\ &= p_X(0,0; \theta) + p_X(1,1; \theta) \\ &= (1 - \theta)^2 + \theta^2. \end{aligned}$$

Note that the incomplete likelihood is convex, with global maxima at $\theta = 0$ and $\theta = 1$, and a global minimum at $\theta = 0.5$. Desired distributions $Q_X \in \mathcal{D}$ take the form

$$q_X(x) = \begin{cases} q_{11} & \text{if } x = (1, 1), \\ 1 - q_{11} & \text{if } x = (0, 0), \\ 0 & \text{otherwise.} \end{cases}$$

The divergence between a desired distribution and a model is given by

$$D(Q_X || P_{X;\theta}) = q_{11} \log \frac{q_{11}}{\theta^2} + (1 - q_{11}) \log \frac{1 - q_{11}}{(1 - \theta)^2},$$

which can be shown to be convex in q_{11} for fixed θ and convex in θ for fixed q_{11} (though not jointly convex in q_{11} and θ). The EM algorithm for estimating θ can be given by a forward step and a backward step as follows:

Forward Step: As described above, the forward step is given by the I-projection of the model $P_{X;\theta^{(t)}}$ onto \mathcal{D} . This is given by

$$q_{X|Y}^{(t+1)}(x|\hat{y}) = p_{X|Y}(x|\hat{y}; \theta^{(t)}),$$

$$q_{11}^{(t+1)} = \frac{\theta^{(t)^2}}{(1 - \theta^{(t)})^2 + \theta^{(t)^2}}.$$

Backward Step: Minimizing the divergence given above over θ for a fixed q_{11} gives

$$\theta^{(t+1)} = q_{11}^{(t+1)}.$$

Thus, the EM iteration for this problem is

$$\theta^{(t+1)} = \frac{\theta^{(t)^2}}{(1 - \theta^{(t)})^2 + \theta^{(t)^2}}.$$

It can be seen that $\theta^{(0)} < 0.5$ gives convergence to the global maximum at $\theta = 0.0$, while $\theta^{(0)} > 0.5$ gives convergence to the global maximum at $\theta = 1.0$. Starting at the global minimum at $\theta = 0.5$ traps the algorithm there.

We now investigate how the additional constraint

$$0.4 \leq q_{11} \leq 0.6 \tag{5}$$

on the desired distribution changes the forward step, and as a result, the convergence behavior of the algorithm. Note that a forward step that projects onto this constrained set of desired distributions will reduce the divergence between the desired distribution and the model, and will therefore be a GAM procedure.

Computing the partial derivative

$$\frac{\partial}{\partial q_{11}} D(Q_X || P_{X;\theta}) = \log \left(\frac{q_{11}}{1 - q_{11}} \left(\frac{1 - \theta}{\theta} \right)^2 \right)$$

shows that it is positive for $0.4 \leq q_{11} \leq 0.6$ when $\theta < \frac{1}{1+\sqrt{3/2}} \approx 0.4495$. Therefore, the forward step from any $\theta < \frac{1}{1+\sqrt{3/2}}$ is given by $q_{11} = 0.4$.

Suppose $\theta^{(0)} = 0.3$. The unconstrained forward step would have given $q_{11}^{(1)} = 0.155$, which would have violated the additional constraint (5). Under the additional constraint (5), the forward step is given by $q_{11}^{(1)} = 0.4$. This in turn leads to $\theta^{(1)} = 0.4$, and the next forward step again gives $q_{11}^{(2)} = 0.4$, showing that the algorithm has converged in a single step, albeit to a value that is not a maximum (or stationary point) in likelihood. Also, recall that the incomplete data likelihood is convex with a minimum at $\theta = 0.5$. This means that initial points in $\theta < 0.4$ will converge in one step to $\theta = 0.4$, thereby reducing likelihood. Indeed, the likelihood at the initial point ($\theta = 0.3$) is 0.58 and the likelihood at the subsequent (limit) points ($\theta = 0.4$) is 0.52.

Thus, it is clear that the convergence behavior of GAM algorithms can differ extremely from that of EM algorithms, and therefore needs to be carefully studied. In fact, non-monotonic convergence behavior can also be seen in the case of incremental EM (Byrne and Gunawardana, 2000). In the following, we will show that under smoothness conditions on the forward and backward steps, GAM procedures that strictly reduce divergence at every step, except possibly at stationary points in likelihood, will yield solutions that are stationary points in likelihood.

3.1 GAM Convergence Theorem

The GAM convergence theorem is a direct application of the generalized convergence theorem (GCT) of Zangwill (1969). We will define the forward and the backward steps to be point-to-set maps, rather than functions, so that we may deal with extended E and M steps that do not yield unique iterates. The GCT will require that these maps be closed. Closedness of a point-to-set map is a smoothness property that is related to function continuity, and is defined as follows:

Definition 1 A point-to-set-map $H : U \rightarrow V$ is closed at $u \in U$ if for any two sequences $\{u^{(t)}\}_{t=0}^\infty \in U$ and $\{v^{(t)}\}_{t=0}^\infty \in V$ the conditions $u^{(t)} \rightarrow u$, $v^{(t)} \rightarrow v$, and $v^{(t)} \in H(u^{(t)})$, imply that $v \in H(u)$.

We now state Zangwill (1969)'s GCT:

Theorem 2 Let the point-to-set map $H : Z \rightarrow Z$ determine an algorithm that given a point $z^{(0)}$ generates a sequence $\{z^{(t)}\}_{t=0}^\infty$ through the iteration $z^{(t+1)} \in H(z^{(t)})$. Also let a solution set Γ be given. Suppose

- (1) All points $z^{(t)}$ are in a compact set $S \subseteq Z$.
- (2) There is a continuous function $\alpha : Z \rightarrow \mathbb{R}$ such that:
 - (a) if $z \notin \Gamma$, then $\alpha(z') < \alpha(z) \forall z' \in H(z)$,
 - (b) if $z \in \Gamma$, then $\alpha(z') \leq \alpha(z) \forall z' \in H(z)$.

- (3) The map H is closed at z if $z \notin \Gamma$.

Then the limit of any convergent subsequence of $\{z^{(t)}\}_{t=0}^\infty$ is in Γ . That is, accumulation points z^* of the sequence $z^{(t)}$ lie in Γ . Furthermore, $\alpha(z^{(t)})$ converges to α^* , and $\alpha(z^*) = \alpha^*$ for all accumulation points z^* .

We use this GCT to show our main convergence result for GAM procedures and then give a corollary that describes how they converge in likelihood.

Theorem 3 (GAM Convergence Theorem) *Let \mathcal{D} be any family of distributions on X and let Θ be the parameter family defined in Section 2. Let the solution set Γ be defined as*

$$\Gamma = \left\{ (Q_X, \theta) : Q_X \in \underset{Q'_X \in \mathcal{D}}{\operatorname{argmin}} D(Q'_X || P_{X;\theta}) \text{ and } \theta \in \underset{\xi \in \Theta}{\operatorname{argmin}} D(Q_X || P_{X;\xi}) \right\}.$$

Let $FB : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$ be any point-to-set map such that all $(Q'_X, \theta') \in FB(Q_X, \theta)$ satisfy

$$(GAM) : \quad D(Q'_X || P_{X;\theta'}) \leq D(Q_X || P_{X;\theta})$$

with equality only if

$$(EQ) : \quad (Q_X, \theta) \in \Gamma.$$

Let $\{(Q_X^{(t)}, \theta^{(t)})\}_{t=0}^{\infty} \in \mathcal{D} \times \Theta$ be a sequence generated from a pair $(Q_X^{(0)}, \theta^{(0)})$ by the iterative application of the point-to-set-map FB :

$$(Q_X^{(t+1)}, \theta^{(t+1)}) \in FB(Q_X^{(t)}, \theta^{(t)}).$$

Suppose that Θ is compact, that there is a compact set $\mathcal{D}' \subseteq \mathcal{D}$ such that

$$(1) \quad FB(\mathcal{D}' \times \Theta) \stackrel{\Delta}{=} \cup_{(Q_X, \theta) \in \mathcal{D}' \times \Theta} FB(Q_X, \theta) \subseteq \mathcal{D}' \times \Theta,$$

(2) The point-to-set map FB is closed on $\mathcal{D}' \times \Theta$,

and that it can be shown that $(Q_X^{(k)}, \theta^{(k)}) \in \mathcal{D}' \times \Theta$ for some iteration (k) .

Then all accumulation points (Q_X^*, θ^*) of the sequence $\{(Q_X^{(t)}, \theta^{(t)})\}_{t=0}^{\infty}$ lie in the solution set Γ and $D(Q_X^* || P_{X;\theta^*}) = D^*$ and $D(Q_X^{(t)} || P_{X;\theta^{(t)}}) \rightarrow D^*$.

Proof We restrict the point-to-set map FB to $\mathcal{D}' \times \Theta$, and then apply Zangwill's GCT above with $S = Z = \mathcal{D}' \times \Theta$, $\alpha = D$, $H = FB$, and $\{z^{(t)}\}_{t=0}^{\infty} = \{(Q_X^{(t)}, \theta^{(t)})\}_{t=k}^{\infty}$. The compactness of $\mathcal{D}' \times \Theta$ follows from the compactness of \mathcal{D}' and Θ individually. The continuity of the divergence in (Q_X, θ) follows from the continuity of the divergence $D(Q_X || P_{X;\theta})$ in Q_X and $P_{X;\theta}$ and the continuity of $p_{X;\theta}$ in θ . The theorem then follows by direct application of Zangwill's theorem. ■

Corollary 4 (Stationary Points in Likelihood) *In Theorem 3, suppose that \mathcal{D} is the desired family defined in Section 2. Then the following hold for accumulation points (Q_X^*, θ^*) :*

$$(1) \quad p_Y(\hat{y}; \theta^*) = e^{-D^*} \text{ and } p_Y(\hat{y}; \theta^{(t)}) \rightarrow e^{-D^*}.$$

(2) θ^* is a stationary point of the incomplete data likelihood if it is in the interior of Θ .

Proof For $(Q_X, \theta) \in \Gamma$, $q_X(x) = p_{X|Y}(x|\hat{y}; \theta)$ so that $D(Q_X || P_{X; \theta}) = -\log q(\hat{y}; \theta)$ yielding conclusion (1).

Since $(Q_X^*, \theta^*) \in \Gamma$, $q_X^*(x) = p_{X|Y}(x|\hat{y}; \theta^*)$, giving

$$\theta^* \in \arg \min_{\theta \in \Theta} D(P_{X|Y=\hat{y}; \theta^*} || P_{X; \theta}).$$

The divergence in the right hand side can be expanded as

$$D(P_{X|Y=\hat{y}; \theta^*} || P_{X; \theta}) = -\log p_Y(\hat{y}; \theta) + D(P_{X|Y=\hat{y}; \theta^*} || P_{X|Y=\hat{y}; \theta}).$$

Taking the gradient of this expression and setting it to zero yields

$$-\nabla_{\theta} \log p_Y(\hat{y}; \theta) \Big|_{\theta=\theta^*} + \nabla_{\theta} D(P_{X|Y=\hat{y}; \theta^*} || P_{X|Y=\hat{y}; \theta}) \Big|_{\theta=\theta^*} = 0.$$

Since $D(P_{X|Y=\hat{y}; \theta^*} || P_{X|Y=\hat{y}; \theta})$ is minimized when $\theta = \theta^*$, this gives us that

$$\nabla_{\theta} \log p_Y(\hat{y}; \theta) \Big|_{\theta=\theta^*} = 0.$$

This proves conclusion (2). ■

The GAM convergence theorem and corollary provide conditions under which iterative estimation procedures converge to stationary points in likelihood. However it is possible that these procedures are not monotonic in likelihood. This can be seen from the Pythagorean equality (Csiszár, 1975) which provides the following relationship between all Q_X in the linear family \mathcal{D} and a model $P_{X; \theta}$

$$D(Q_X || P_{X; \theta}) = D(Q_X || \tilde{Q}_X) + D(\tilde{Q}_X || P_{X; \theta})$$

where the I-projection $\tilde{Q}_X = \operatorname{argmin}_{Q_X \in \mathcal{D}} D(Q_X || P_{X; \theta})$ is uniquely specified as $\tilde{Q}_{X|Y=\hat{y}} = P_{X|Y=\hat{y}; \theta}$. From this we find the following relationship between the likelihood of the model estimates and the overall divergence

$$D(Q_X || P_{X; \theta}) = D(Q_X || P_{X|Y=\hat{y}; \theta}) - \log p_Y(\hat{y}; \theta).$$

While GAM procedures guarantee that $D(Q_X^{(t+1)} || P_{X; \theta^{(t+1)}}) \leq D(Q_X^{(t)} || P_{X; \theta^{(t)}})$, we can conclude only that

$$\log p_Y(\hat{y}; \theta^{(t+1)}) \geq \log p_Y(\hat{y}; \theta^{(t)}) + \Delta^{(t)}$$

where $\Delta^{(t)} = D(Q_X^{(t+1)} || P_{X|Y=\hat{y}; \theta^{(t+1)}}) - D(Q_X^{(t)} || P_{X|Y=\hat{y}; \theta^{(t)}})$. Since, as shown in Figure 3, this quantity can be negative, it is possible for GAM algorithms to be non-monotonic in likelihood even while converging to local maxima in likelihood.

We now discuss the construction of a GAM mapping FB that satisfies the requirements of the GAM convergence theorem.

Proposition 5 *Let the point-to-set map FB in Theorem 3 above be the composition $B \circ F$ of point-to-set maps $F : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$ and $B : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$. Suppose that the point-to-set maps F and B are defined so that*

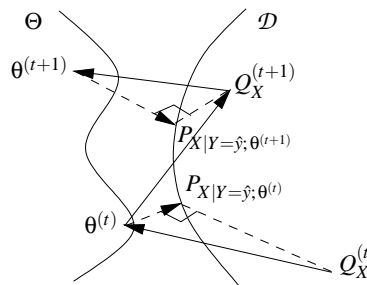


Figure 3: A schematic representation of how GAM procedures may be non-monotonic in likelihood. The solid arrows show forward and backward steps that reduce the divergence rather than minimizing it. The broken arrows show the forward steps that would have been taken by the EM algorithm (i.e., the I-projections of the models). Divergences that obey the Pythagorean equality are indicated by right triangles. In particular, the squared lengths of the broken arrows represent negative log likelihood. Note that the divergence between the desired distribution yielded by the forward step and the I-projection of the model decreases, while the negative log likelihood increases.

(1) F and B are closed on $\mathcal{D}' \times \Theta$

(2) $F(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$ and $B(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$

Suppose also that F is such that all $(Q'_X, \theta') \in F(Q_X, \theta)$ have $\theta' = \theta$ and satisfy

$$(GAM.F) : \quad D(Q'_X || P_{X;\theta}) \leq D(Q_X || P_{X;\theta})$$

with equality only if

$$(EQ.F) : \quad Q_X = \operatorname{argmin}_{Q'_X \in \mathcal{D}} D(Q'_X || P_{X;\theta}),$$

with Q_X being the unique minimizer. Suppose also that the point-to-set map B is such that all $(Q'_X, \theta') \in B(Q_X, \theta)$ have $Q'_X = Q_X$ and satisfy

$$(GAM.B) : \quad D(Q_X || P_{X;\theta'}) \leq D(Q_X || P_{X;\theta})$$

with equality only if

$$(EQ.B) : \quad \theta \in \operatorname{argmin}_{\xi \in \Theta} D(Q_X || P_{X;\xi}).$$

Then,

(1) the point-to-set map FB is closed on $\mathcal{D}' \times \Theta$

(2) $FB(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$

and FB satisfies the GAM and EQ conditions of the GAM convergence theorem.

Proof If the point-to-set maps $F : A \rightarrow B$ and $G : B \rightarrow C$ are closed on A and B respectively, their composition $FG = G \circ F$ is closed on A if B is compact. Since F and B are closed on $\mathcal{D}' \times \Theta$, which is compact, it follows that FB is closed on $\mathcal{D}' \times \Theta$. That $FB(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$ follows directly from the assumptions of the proposition.

The condition (GAM) follows directly from (GAM.F) and (GAM.B).

Conditions (EQ.F) and (EQ.B) together are not enough to ensure condition (EQ). Suppose $(R_X, \phi) \in FB(Q_X, \theta)$. This implies that $(R_X, \theta) \in F(Q_X, \theta)$ and $(R_X, \phi) \in B(R_X, \theta)$.

Suppose $D(R_X || P_{X;\phi}) = D(Q_X || P_{X;\theta})$. Then (GAM.F) and (GAM.B) ensure that $D(R_X || P_{X;\phi}) = D(R_X || P_{X;\theta}) = D(Q_X || P_{X;\theta})$. Condition (EQ.F) gives

$$\begin{aligned} Q_X &\in \arg \min_{Q'_X \in \mathcal{D}} D(Q'_X || P_{X;\theta}), \\ R_X &\in \arg \min_{Q'_X \in \mathcal{D}} D(Q'_X || P_{X;\theta}), \end{aligned} \tag{6}$$

and (EQ.B) gives

$$\theta \in \arg \min_{\xi \in \Theta} D(R_X || P_{X;\xi}). \tag{7}$$

While equation (6) is the first criterion for membership in Γ , equation (7) is not quite the second criterion – the divergence minimized here is $D(R_X || P_{X;\xi})$ instead of $D(Q_X || P_{X;\xi})$. Since by assumption, Q_X is the unique minimizer of the divergence, $Q_X = R_X$, giving the required condition

$$\theta \in \arg \min_{\xi \in \Theta} D(Q_X || P_{X;\xi}).$$

■

This allows us to construct a map FB through the composition of generalized forward and backward steps F and B . As seen in the proof it is insufficient for the forward and backward steps to satisfy the GAM and EQ conditions separately. It is also necessary for the forward step to satisfy the equality condition with a unique minimizer. For example, this condition is satisfied when \mathcal{D} is defined by linear constraints as in Section 2 and the forward step is a simple projection, as in the case of EM. Even when this condition is not satisfied, it may be possible to show condition (EQ) for the composite map FB . It is important to show that FB strictly decreases the divergence for all points outside the solution set Γ , since any points where this does not hold are accumulation points of the algorithm.

As an instance of the GAM procedure, EM convergence is also explained by these results as shown in Appendix A. The conditions of Theorem 3 and Corollary 4 are quite general, and very similar to those that must be satisfied to ensure GEM convergence (Wu, 1983). For example, in both GEM and GAM, condition (Q2) must hold. Insisting on this would rule out GMMs with parameter families that allow individual Gaussians to have a variance of zero. In practice, modeling considerations usually prevent such situations.

4. Incremental EM as GAM

We now turn our attention to the incremental EM algorithm of Neal and Hinton (1998). This variant of the EM algorithm divides the training data into partitions, and at each iterate, computes conditional sufficient statistics on only one partition. The statistics conditioned on other partitions are

saved from previous iterations. The statistics corresponding to the different partitions are pooled before performing the M step at each iteration, but the separate per-partition statistics are retained for use in future iterations. This algorithm has shown to give faster convergence in a number of applications (Digalakis, 1997; Thiesson et al., 2001; Hsiao et al., 2004), even though it may be non-monotonic in likelihood (Byrne and Gunawardana, 2000). Here, we use our GAM results to show that in most cases, the incremental updates do not sacrifice the convergence guarantees of EM, despite the non-monotonicity in likelihood. Note that Neal and Hinton (1998) have shown that incremental EM is monotonic in divergence, but not that it converges to EM fixed points.

The complete variable $X = (X^{(1)}, \dots, X^{(n)})$ is assumed to consist of n independent components so that $Q_X = \prod_{i=1}^n Q_{X^{(i)}}$. The visible variable $Y = (Y^{(1)}, \dots, Y^{(n)})$ has observed value $\hat{y} = (\hat{y}^{(1)}, \dots, \hat{y}^{(n)})$. The components $Y^{(i)}$ are generated independently of each other, from their corresponding $X^{(i)}$.

The EM auxiliary function for these variables is

$$\begin{aligned} \Phi(\theta|\theta^{(t)}) &= \sum_{i=1}^n \mathbf{E}_{P_{X^{(i)}|Y^{(i)}}} \left[\log p_{X^{(i)}}(X^{(i)}; \theta) \mid \hat{y}^{(i)}; \theta^{(t)} \right] \\ &= \sum_{i=1}^n \Phi^{(i)}(\theta|\theta^{(t)}). \end{aligned}$$

Rather than maximize this auxiliary function, the incremental EM algorithm allows re-estimation to be performed based on a single component $\hat{y}^{(i)}$ of the observation \hat{y} at any step. For example, in a two-element problem the re-estimation procedure might proceed as follows :

$$\begin{aligned} \theta^{(t+1)} &= \operatorname{argmax}_{\theta \in \Theta} (\Phi^{(1)}(\theta|\theta^{(t-1)}) + \Phi^{(2)}(\theta|\theta^{(t)})), \\ \theta^{(t+2)} &= \operatorname{argmax}_{\theta \in \Theta} (\Phi^{(1)}(\theta|\theta^{(t+1)}) + \Phi^{(2)}(\theta|\theta^{(t)})), \\ \theta^{(t+3)} &= \operatorname{argmax}_{\theta \in \Theta} (\Phi^{(1)}(\theta|\theta^{(t+1)}) + \Phi^{(2)}(\theta|\theta^{(t+2)})), \\ &\dots \end{aligned}$$

This is not enough to ensure that $\Phi(\theta^{(t+3)}|\theta^{(t+1)}) \leq \Phi(\theta^{(t+1)}|\theta^{(t+1)})$ so the (G)EM convergence results do not apply. However the algorithm can be formulated as an GAM procedure.

To show that incremental EM can be a GAM procedure, we describe it as a nested series of n incremental forward steps and n exact backward steps. Iteration $(t+1)$ of incremental EM proceeds as follows. First, the iteration is initialized from the results of the previous iteration:

$$Q_X^{(t+1,0)} = Q_X^{(t)} \text{ and } \theta^{(t+1,0)} = \theta^{(t)}.$$

We then define a series of n incremental forward steps $j = 1, \dots, n$

$$Q_{X^{(i)}}^{(t+1,j)} = \begin{cases} P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \theta^{(t+1,j-1)}} & \text{if } j = i \\ Q_{X^{(i)}}^{(t+1,j-1)} & \text{otherwise,} \end{cases}$$

and backward steps

$$\theta^{(t+1,j)} \in \operatorname{argmin}_{\xi \in \Theta} D(Q_X^{(t+1,j)} \| P_{X;\xi}),$$

so that finally we set $Q_X^{(t+1)} = Q_X^{(t+1,n)}$ and $\theta^{(t+1)} = \theta^{(t+1,n)}$.

We formally represent the (j) th incremental forward step $F^{(j)} : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$ as the singleton point-to-set map

$$F^{(j)}(Q_X, \theta) = \left\{ (Q'_X, \theta') : Q'_X = P_{X^{(j)}|Y^{(j)}=\hat{y}^{(j)}; \theta} \prod_{i \neq j} Q_{X^{(i)}} \right\}.$$

It updates the (j) th component marginal $Q_{X^{(j)}}$ of Q_X but keeps the other component marginals fixed.

The backward step is represented by a closed point-to-set map $B : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$ satisfying conditions (GAM.B) and (EQ.B) of Proposition 5 with $Q'_X = Q_X$ for $(Q'_X, \theta') \in B(Q_X, \theta)$, and additionally satisfying:

$$B(Q_X, \theta) \text{ is a singleton set } \forall (Q_X, \theta) \in \mathcal{D} \times \Theta : (Q_X, \theta) \in M(Q_X, \theta). \quad (8)$$

Thus, we are guaranteed that $\theta' = \theta$ when $D(Q_X || P_{X; \theta'}) = D(Q_X || P_{X; \theta})$. This is equivalent to requiring that the EM auxiliary function has a unique maximizer. We note that this often holds in practice – for example, when the complete data distribution comes from a flat exponential family (Efron, 1975; Amari, 1995) as is the case with mixtures of Gaussians, or with hidden Markov models. Even when the complete data distribution is a curved exponential family, uniqueness can still be possible.

Using these composite maps we can describe incremental EM as

$$(Q_X^{(t+1)}, \theta^{(t+1)}) \in FB(Q_X^{(t)}, \theta^{(t)})$$

where

$$FB = B \circ F^{(n)} \circ \dots \circ B \circ F^{(1)}.$$

Proposition 6 *As defined, incremental EM can be shown to converge to stationary points in likelihood through application of the GAM convergence theorem.*

Proof For any $(Q_X, \theta) \in \mathcal{D} \times \Theta$, we use the independence of the components $X^{(i)}$ and $Y^{(i)}$ to decompose the divergence $D(Q_X || P_{X; \theta})$ into a sum of component divergences as follows

$$D(Q_X || P_{X; \theta}) = \sum_i D(Q_{X^{(i)}} || P_{X^{(i)}; \theta}).$$

The (j) th backward step satisfies

$$\begin{aligned} D(Q_X^{(t+1,j)} || P_{X; \theta^{(t+1,j)}}) &\leq D(Q_X^{(t+1,j)} || P_{X; \theta^{(t+1,j-1)}}) \\ &= \sum_{i: i \neq j} D(Q_{X^{(i)}}^{(t+1,j-1)} || P_{X; \theta^{(t+1,j-1)}}) + D(Q_{X^{(j)}}^{(t+1,j)} || P_{X; \theta^{(t+1,j-1)}}) \end{aligned}$$

where the right hand side has been expanded using the fact that the (j) th incremental forward step leaves all but the (j) th component divergence unchanged. Since the (j) th incremental forward step minimizes the (j) th component divergence, we get

$$\begin{aligned} D(Q_X^{(t+1,j)} || P_{X; \theta^{(t+1,j)}}) &\leq \sum_{i: i \neq j} D(Q_{X^{(i)}}^{(t+1,j-1)} || P_{X; \theta^{(t+1,j-1)}}) + D(Q_{X^{(j)}}^{(t+1,j-1)} || P_{X; \theta^{(t+1,j-1)}}) \\ &= D(Q_X^{(t+1,j-1)} || P_{X; \theta^{(t+1,j-1)}}). \end{aligned}$$

Condition (GAM) of Theorem 3 is therefore satisfied.

Since the maps $F^{(j)}$ and B are closed (Appendix A, Proposition 7), FB will be closed on \mathcal{D}' , if the set is constructed so as to be compact (Appendix A). An appropriate definition of $\mathcal{D}' \subseteq \mathcal{D}$ is given in Appendix B along with a proof that incremental EM satisfies the equality condition (EQ) of Theorem 3. ■

Thus, the GAM convergence theorem shows that incremental EM procedures converge to EM fixed points when the EM auxiliary function is uniquely maximized. However, it is not a GEM procedure, and monotonicity in likelihood is no longer guaranteed. Indeed, as discussed in Byrne and Gunawardana (2000) non-monotonicity in likelihood is observed in practice, and the convergence behavior is very different from that of (G)EM procedures, despite the common fixed point set. Thiesson et al. (2001) also show that the convergence behavior of incremental EM is different from that of EM in practice.

5. Variational EM as GAM

Variational approximations have been popular in cases where computing the exact forward step $q_X(x) = p_{X|Y}(x|\hat{y}; \theta)$ is intractable (Jordan et al., 1999). The idea is to restrict attention to a subfamily \mathcal{D}_V of \mathcal{D} such that members of \mathcal{D}_V have a particular parametric form, which is chosen so that projecting a model $P_{X; \theta}$ onto \mathcal{D}_V is more tractable than projecting it onto \mathcal{D} . That is, a parametrization $q_X(x; \lambda)$ with $\lambda \in \Lambda$ is fixed, and the family \mathcal{D}_V is defined as

$$\mathcal{D}_V = \{Q_X \in \mathcal{D} : q_X(x) = q_X(x; \lambda) \text{ for some } \lambda \in \Lambda\}.$$

We assume $\Lambda \subseteq \mathbb{R}^n$ is closed and bounded.

Then, the variational forward step is defined to be

$$F_V(Q_X, \theta) = \left\{ (Q'_X, \theta) : Q'_X \in \arg \min_{Q''_X \in \mathcal{D}_V} D(Q''_X \| P_{X; \theta}) \right\}.$$

By the Pythagorean equality of Csiszár (1975),

$$\begin{aligned} D(Q''_X \| P_{X; \theta}) &= D(Q''_X \| P_{X|Y=\hat{y}; \theta}) + D(P_{X|Y=\hat{y}; \theta} \| P_{X; \theta}) \\ &= D(Q''_X \| P_{X|Y=\hat{y}; \theta}) - \log p_Y(\hat{y}; \theta). \end{aligned}$$

Thus, $Q''_X \in \mathcal{D}_V$ that minimizes this divergence also best approximates $P_{X|Y=\hat{y}; \theta}$, which is the desired distribution that would be chosen by the usual EM procedure.

Notice that the divergence minimized at every iteration is no longer just $D(P_{X|Y=\hat{y}; \theta} \| P_{X; \theta})$ (which is the negative log likelihood) as in the EM algorithm, and that therefore, the likelihood is not guaranteed to increase at every iteration. We now examine if the conditions of the GAM convergence theorem of Section 3 still hold if the forward step of the EM procedure is replaced by F_V .

First, note that \mathcal{D}_V is a natural choice for \mathcal{D}' as long as the set of variational parameters Λ is compact. That the map F_V is closed on $\mathcal{D}_V \times \Theta$ follows from Corollary 8 and Lemma 9 of Appendix A, and the assumptions on Λ . The mapping satisfies conditions (GAM.F) and (EQ.F) because each new desired distribution must minimize the divergence to \mathcal{D}_V . However, the uniqueness condition

of Proposition 5 (EQ.F) cannot be guaranteed in general, and must be verified for each choice of \mathcal{D}_V . If this condition holds, then the algorithm converges to minimizers of the divergence between the family of variational approximations and the model family. For example, this happens when the variational E step is uniquely defined.

We now analyze when these limit points (Q_X^*, θ^*) are stationary points in likelihood. Since θ^* minimizes the divergence $D(Q^* || P_{X; \theta})$ over θ ,

$$\nabla_{\theta} D(Q_X^* || P_{X; \theta}) \Big|_{\theta=\theta^*} = 0.$$

Expanding the divergence as before,

$$\nabla_{\theta} D(Q_X^* || P_{X|Y=\hat{y}; \theta}) \Big|_{\theta=\theta^*} - \nabla_{\theta} \log q(\hat{y}; \theta) \Big|_{\theta=\theta^*} = 0$$

so that $\nabla_{\theta} \log q(\hat{y}; \theta) \Big|_{\theta=\theta^*} = 0$ if and only if

$$\nabla_{\theta} D(Q_X^* || P_{X|Y=\hat{y}; \theta}) \Big|_{\theta=\theta^*} = 0.$$

Therefore a θ^* generated by a variational EM procedure is a stationary point in likelihood if and only if θ^* is a parameter that locally minimizes the variational approximation error. This can happen in two ways. First, the variational error may have stationary points at stationary points in likelihood. This can only be ensured if the stationary points are known before estimation. Second, the variational error is independent of θ . This is not possible if the variational family introduces independence assumptions that ensure tractability. In particular, a model which agrees with the variational approximation (e.g., a factorial HMM with parameter settings that decouple the state sequences) will have lower variational error than one that does not. We illustrate this in the case of the mean field approximation for Boltzmann machines.

Example 3 *In Example 1, choose a pair of hidden nodes i, j connected by a dependency link. It is well-known (Byrne, 1992) that*

$$\begin{aligned} \frac{\partial}{\partial \theta_{ij}} \log P_{S|S_E}(s|\hat{s}_E; \theta) &= s_i s_j - \mathbf{E}_{P_{S|S_E}} [S_i S_j | \hat{s}_E; \theta], \\ \frac{\partial}{\partial \theta_{k0}} \log P_{S|S_E}(s|\hat{s}_E; \theta) &= s_k - \mathbf{E}_{P_{S|S_E}} [S_k | \hat{s}_E; \theta] \quad k = i, j, \end{aligned}$$

which gives

$$\begin{aligned} \frac{\partial}{\partial \theta_{ij}} D(Q_S^* || P_{S|S_E=\hat{s}_E; \theta}) &= \mathbf{E}_{Q_S^*} [S_i S_j; \mu] - \mathbf{E}_{P_{S|S_E}} [S_i S_j | \hat{s}_E; \theta] \\ &= \mu_i^* \mu_j^* - \mathbf{E}_{P_{S|S_E}} [S_i S_j | \hat{s}_E; \theta], \\ \frac{\partial}{\partial \theta_k} D(Q_S^* || P_{S|S_E=\hat{s}_E; \theta}) &= \mu_k^* - \mathbf{E}_{P_{S|S_E}} [S_k | \hat{s}_E; \theta] \quad k = i, j. \end{aligned}$$

If $\nabla_{\theta} D(Q_S^* || P_{S|S_E=\hat{s}_E;\theta}) \Big|_{\theta=\theta^*}$ is to be zero,

$$\mathbf{E}_{P_{S|S_E}} [S_i S_j | \hat{s}_E; \theta^*] = \mathbf{E}_{P_{S|S_E}} [S_i | \hat{s}_E; \theta^*] \mathbf{E}_{P_{S|S_E}} [S_j | \hat{s}_E; \theta^*]$$

must hold. This can only occur if $\theta_{ij}^* = 0$, when the model itself satisfies the constraints of the mean field approximation. Thus, under this variational approximation, only Boltzmann machines where the hidden units do not depend on each other can give stationary points in likelihood.

To summarize, the GAM convergence theorem applies to variational EM in cases when the variational E step is uniquely defined. When this is so, the resulting model is at a local minimum of the divergence between the model family and the family of variational approximations. However, except in degenerate cases, this model cannot be at a stationary point in likelihood. These degenerate cases occur when the model satisfies the simplifying conditions that define the family of variational approximations. In this case, the variational EM algorithm is essentially performing standard EM over a restricted model family defined so as to be consistent with the variational approximations.

6. Conclusion

GAM iterative estimation procedures are a class of EM extensions whose E step can be varied in a manner analogous to the relaxation of the M step that occurs in GEM algorithms. We have provided conditions under which these procedures can be shown to converge to stationary points in likelihood. The conditions specify allowable E step variations that are in fact analogous to the M step variations that are allowed by GEM procedures. The convergence analysis is analogous to that presented by Wu (1983), but takes advantage of the information geometric framework of Csiszár and Tusnády (1984) to explicitly represent distributions in computing sufficient statistics.

We have analyzed the convergence behavior of two well known EM extensions, namely incremental EM and variational EM, as GAM procedures. Our GAM convergence analysis shows that incremental EM procedures converge to stationary points in likelihood, even though incremental EM is in general neither a GEM procedure nor monotonic in likelihood. Variational EM algorithms with unique E steps satisfy the conditions of the GAM convergence theorem but do not satisfy its corollary. Thus the GAM convergence theorem shows that such algorithms converge to solutions that minimize divergence, but these are not necessarily stationary points in likelihood. We then present an information geometric argument which shows that variational EM can only converge to stationary points in likelihood in degenerate cases.

Appendix A. EM Satisfies the GAM Convergence Theorem

Recall that the forward and backward steps $F, B : \mathcal{D} \times \Theta \rightarrow \mathcal{D} \times \Theta$ of the EM algorithm are given by

$$F(Q_X, \theta) = \left\{ (Q'_X, \theta) : Q'_X \in \arg \min_{Q'_X \in \mathcal{D}} D(Q'_X || P_{X;\theta}) \right\}$$

and

$$B(Q_X, \theta) = \left\{ (Q_X, \phi) : \phi \in \arg \min_{\xi \in \Theta} D(Q_X || P_{X;\xi}) \right\}.$$

That the conditions (GAM.F), (GAM.B), (EQ.F), and (EQ.B) hold is obvious by construction. We will first show a compact \mathcal{D}' that guarantees that $F(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$ and $B(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$. We will then show that F and B are closed on $\mathcal{D}' \times \Theta$. Proposition 5 then implies that the composite map FB satisfies the conditions of the GAM convergence theorem (Theorem 3).

Restricting the desired family to a compact set We define \mathcal{D}' as

$$\mathcal{D}' = \{Q_X \in \mathcal{D} : q_{X|Y}(x|\hat{y}) = p_{X|Y}(x|\hat{y}; \theta) \text{ for some } \theta \in \Theta\}$$

with \mathcal{D} defined as in Section 2. Note that this forces every $Q_X \in \mathcal{D}'$ to be the continuous mapping of some $\theta \in \Theta$. Therefore, \mathcal{D}' is the continuous mapping of the compact set Θ , and is therefore compact.

By construction of \mathcal{D}' , it is guaranteed that Q_X generated by a forward step will lie in \mathcal{D}' . Thus, $F(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$. By definition of $B(\cdot)$, $B(\mathcal{D}' \times \Theta) \subseteq \mathcal{D}' \times \Theta$.

Closedness of the forward and backward steps The following proposition and corollary show that the minimization of a continuous function forms a closed point-to-set map. This implies that projection under the divergence forms a closed point-to-set-map, so that the (ungeneralized) forward and backward steps of the EM algorithm are in fact closed point-to-set maps.

Proposition 7 *Given a real-valued continuous function f on $A \times B$, define the point-to-set map $F : A \rightarrow B$ by*

$$\begin{aligned} F(a) &= \arg \min_{b' \in B} f(a, b'), \\ &= \{b : f(a, b) \leq f(a, b') \text{ for } \forall b' \in B\}. \end{aligned}$$

Then, the point-to-set map F is closed at a if $F(a)$ is nonempty.

Proof Let $\{a^{(t)}\}_{t=0}^\infty$ and $\{b^{(t)}\}_{t=0}^\infty$ be sequences in A and B respectively, such that

$$\begin{aligned} a^{(t)} &\rightarrow a, \\ b^{(t)} &\rightarrow b, \end{aligned}$$

and suppose

$$b^{(t)} \in F\left(a^{(t)}\right).$$

That is,

$$b^{(t)} \in \arg \min_{b' \in B} f(a^{(t)}, b').$$

The map F is closed at $a \in A$ if this implies that $b \in F(a)$ – that is, that $b \in \arg \min_{b' \in B} f(a, b')$.

To prove the proposition by contradiction, suppose $b \notin \arg \min_{b' \in B} f(a, b')$. By assumption $F(a)$ is nonempty. Therefore, there exists $\hat{b} \in \arg \min_{b' \in B} f(a, b')$. Choose $\varepsilon > 0$ such that

$$f(a, b) > f(a, \hat{b}) + 2\varepsilon. \tag{9}$$

By continuity of $f(\cdot, \cdot)$ and f -monotonicity of $(a^{(t)}, b^{(t)})$, $\exists K_1$ such that

$$f(a^{(t)}, b^{(t)}) > f(a, b) - \varepsilon, \quad \forall t > K_1,$$

so that by equation (9),

$$f(a^{(t)}, b^{(t)}) > f(a, \hat{b}) + \varepsilon, \quad \forall t > K_1.$$

By continuity of $f(\cdot, \hat{b})$ and f -monotonicity of $(a^{(t)}, b^{(t)})$, $\exists K_2$ such that

$$f(a, \hat{b}) + \varepsilon > f(a^{(t)}, \hat{b}), \quad \forall t > K_2.$$

Combining these two bounds gives $\exists t > K_1, K_2$ such that

$$f(a^{(t)}, b^{(t)}) > f(a^{(t)}, \hat{b})$$

which is a contradiction since by assumption, $b^{(t)} \in \arg \min_{b' \in B} f(a^{(t)}, b')$, and therefore,

$$b \in \arg \min_{b' \in B} f(a, b'),$$

$b \in F(a)$. ■

Corollary 8 *The point-to-set map $F : A \rightarrow B$ of Proposition 7 is closed on A if the set B is closed.*

The following lemma shows that the Cartesian product of two closed point-to-set-maps is itself closed.

Lemma 9 *Suppose $F : A \rightarrow B$ and $G : A \rightarrow C$ are closed point-to-set-maps. Then the product point-to-set-map $H : A \rightarrow B \times C$ defined by*

$$H(a) = F(a) \times G(a)$$

is closed.

This follows by direct application of the definition of closedness of point-to-set maps.

Proposition 7 and the existence of the I-projection shows that the mapping from Θ to \mathcal{D} defined by

$$Q_X \in \arg \min_{Q'_X \in \mathcal{D}} D(Q'_X || P_{X;\theta})$$

is closed. This result together with Lemma 9 then show that the forward step F of the EM algorithm shown above is closed. Similarly, it can be shown using Corollary 8 and Lemma 9 that the backward step B is closed.

Appendix B. Incremental EM: (EQ) and \mathcal{D}'

In this appendix, we show that incremental EM satisfies condition EQ of the GAM convergence theorem, and show a compact restriction of the desired family that can be used to analyze convergence of incremental EM.

B.1 Incremental EM Satisfies Condition (EQ)

$(Q'_X, \theta') \in FB(Q_X, \theta)$ implies a sequence of incremental steps

$$(R_X^{(0)}, \phi^{(0)}), \dots, (R_X^{(n)}, \phi^{(n)})$$

such that

$$\begin{aligned} (R_X^{(0)}, \phi^{(0)}) &= (Q_X, \theta), \\ (R_X^{(j)}, \phi^{(j)}) &\in F^{(j)}B(R_X^{(j-1)}, \phi^{(j-1)}), \end{aligned}$$

and

$$(Q'_X, \theta') = (R_X^{(n)}, \phi^{(n)}).$$

When $D(Q'_X || P_{X;\theta'}) = D(Q_X || P_{X;\theta})$, the GAM inequality (already shown) gives that the divergence is unchanged at every incremental step:

$$D(R_X^{(j)} || P_{X;\phi^{(j)}}) = D(R_X^{(j-1)} || P_{X;\phi^{(j-1)}})$$

for $j = 1, \dots, n$. In fact, by conditions (GAM.F) and (GAM.B), the divergence is unchanged at each incremental forward step $F^{(j)}$ and the backward step B :

$$D(R_X^{(j)} || P_{X;\phi^{(j-1)}}) = D(R_X^{(j-1)} || P_{X;\phi^{(j-1)}}) \tag{10}$$

and

$$D(R_X^{(j)} || P_{X;\phi^{(j)}}) = D(R_X^{(j)} || P_{X;\phi^{(j-1)}}) \tag{11}$$

for $j = 1, \dots, n$. We will now show that $Q_{X^{(i)}} = P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \phi^{(i-1)}}$ for $i = 1, \dots, n$ and that $\phi^{(j)} = \theta$ for $j = 1, \dots, n$, which will then imply that condition (EQ) holds.

To show $Q_{X^{(i)}} = P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \phi^{(i-1)}}$ we decompose equation (10) as

$$\begin{aligned} \sum_{i \neq j} D(R_{X^{(i)}}^{(j)} || P_{X;\phi^{(j-1)}}) + D(R_{X^{(j)}}^{(j)} || P_{X;\phi^{(j-1)}}) = \\ \sum_{i \neq j} D(R_{X^{(i)}}^{(j-1)} || P_{X;\phi^{(j-1)}}) + D(R_{X^{(j)}}^{(j-1)} || P_{X;\phi^{(j-1)}}). \end{aligned}$$

Since $R_{X^{(i)}}^{(j)} = R_{X^{(i)}}^{(j-1)}$ for all $i \neq j$ at any incremental step (j) , this reduces to

$$D(R_{X^{(j)}}^{(j)} || P_{X;\phi^{(j-1)}}) = D(R_{X^{(j)}}^{(j-1)} || P_{X;\phi^{(j-1)}})$$

for $j = 1, \dots, n$. Since $R_{X^{(j)}}^{(j)} = P_{X^{(j)}|Y^{(j)}=\hat{y}^{(j)}; \phi^{(j-1)}}$ uniquely minimizes the component divergence $D(R_{X^{(j)}} || P_{X; \phi^{(j-1)}})$ over all $R_{X^{(j)}}$, this means that

$$R_{X^{(j)}}^{(j-1)} = R_{X^{(j)}}^{(j)} = P_{X^{(j)}|Y^{(j)}=\hat{y}^{(j)}; \phi^{(j-1)}}.$$

Thus, for any component (i) , substituting in $j = i$ and recalling that the first $i - 1$ incremental E steps leave the component marginal $R_{X^{(i)}}$ unchanged, we get

$$Q_{X^{(i)}} = R_{X^{(i)}}^{(0)} = \dots = R_{X^{(i)}}^{(i)} = P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \phi^{(i-1)}}. \quad (12)$$

We now show that $\phi^{(j)} = \theta$ for $j = 1, \dots, n$. Since equation (11) tells us that the divergence is unchanged at any backward step (j) , both $P_{X; \phi^{(j)}}$ and $P_{X; \phi^{(j-1)}}$ must minimize $D(R_X^{(j)} || \cdot)$. By assumption (8), the M-step is uniquely determined, so we must have $\phi^{(j)} = \phi^{(j-1)}$. We therefore have the desired result

$$\theta = \phi^{(0)} = \dots = \phi^{(n)} = \theta'.$$

Substituting this into equation (12) gives

$$Q_{X^{(i)}} = R_{X^{(i)}}^{(0)} = \dots = R_{X^{(i)}}^{(i)} = P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \theta}.$$

Since this applies for all $i = 1, \dots, n$, we have

$$R_X^{(j)} = P_{X|Y=\hat{y}; \theta}, \quad \forall j = 0, \dots, n.$$

In particular, $Q_X = R_X^{(0)} = P_{X|Y=\hat{y}; \theta}$, which means that

$$Q_X = \operatorname{argmin}_{Q_X'' \in \mathcal{D}} D(Q_X'' || P_{X; \theta}).$$

Since $R_X^{(1)} = Q_X$, we use equation (11) with $j = 1$ and condition (EQ.B) on the backward map to get

$$\theta \in \operatorname{argmin}_{\xi \in \Theta} D(Q_X || P_{X; \xi}).$$

This shows that condition (EQ) holds.

B.2 Definition of a Compact \mathcal{D}'

To find a suitable restriction \mathcal{D}' for any choice of $Q_X^{(0)} \in \mathcal{D}$, we first define the following sets of measures on the components $X^{(i)}$:

$$\mathcal{D}_{INC}^{(i)} = \left\{ Q_{X^{(i)}} : Q_{X^{(i)}} = P_{X^{(i)}|Y^{(i)}=\hat{y}^{(i)}; \theta} \text{ for some } \theta \in \Theta \right\} \cup \left\{ Q_{X^{(i)}}^{(0)} \right\},$$

and note that the continuity of $P_{X|Y; \theta}$ (assumed), and the compactness of Θ (assumed) give us compactness of $\mathcal{D}_{INC}^{(i)}$. We then define our restriction \mathcal{D}'_{INC} of \mathcal{D} by

$$\mathcal{D}'_{INC} = \left\{ Q_X : Q_X = \prod_{i=1}^n Q_{X^{(i)}} \text{ for some } (Q_{X^{(1)}}, \dots, Q_{X^{(n)}}) \in \mathcal{D}_{INC}^{(1)} \times \dots \times \mathcal{D}_{INC}^{(n)} \right\}.$$

To show compactness of \mathcal{D}'_{INC} , suppose $\{Q_X^{(t)}\}_{t=0}^{\infty}$ is a sequence in \mathcal{D}'_{INC} . From the definition of \mathcal{D}'_{INC} , this then implies that there are n sequences $\{Q_{X^{(i)}}^{(t)}\}_{t=0}^{\infty}$, each in the corresponding $\mathcal{D}_{INC}^{(i)}$, such that $Q_X^{(t)} = \prod_{i=1}^n Q_{X^{(i)}}^{(t)}$. The compactness of $\mathcal{D}_{INC}^{(1)}$ implies the existence of an infinite subset $\mathcal{K}^{(1)}$ of the integers such that the subsequence $\{Q_{X^{(1)}}^{(t)}\}_{t \in \mathcal{K}^{(1)}}$ converges to some $Q_{X^{(1)}}^* \in \mathcal{D}_{INC}^{(1)}$. Similarly, since the infinite sequence $\{Q_{X^{(i)}}^{(t)}\}_{t \in \mathcal{K}^{(i-1)}}$ is contained in the compact set $\mathcal{D}_{INC}^{(i)}$, there exists an infinite subset $\mathcal{K}^{(i)}$ of $\mathcal{K}^{(i-1)}$ such that the subsequence $\{Q_{X^{(i)}}^{(t)}\}_{t \in \mathcal{K}^{(i)}}$ converges to some $Q_{X^{(i)}}^* \in \mathcal{D}_{INC}^{(i)}$. Therefore, the subsequence $\{Q_X^{(t)}\}_{t \in \mathcal{K}^{(n)}}$ converges to $\prod_{i=1}^n Q_{X^{(i)}}^* \in \mathcal{D}'_{INC}$, showing that \mathcal{D}'_{INC} is compact.

References

- S.-I. Amari. Information geometry of the EM and *em* algorithms for neural networks. *Neural Networks*, 8(9):1379–1408, 1995.
- R. A. Boyles. On the convergence of the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 45(1):47–50, 1983.
- W. Byrne. Alternating minimization and Boltzman machine learning. *IEEE Transactions on Neural Networks*, 3(4):612–620, 1992.
- W. Byrne and A. Gunawardana. Comments on ‘Efficient training algorithms for HMM’s using incremental estimation’. *IEEE Transactions on Speech and Audio Processing*, 8(6):751–754, November 2000.
- I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, 1975.
- I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplemental Issue Number 1*, pages 205–237, 1984.
- I. Csiszár. Information theory and statistics. ENEE 728F. Lecture Notes, Spring 1990. University of Maryland.
- J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.
- A. P. Dempster, A. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.
- V. Digalakis. On-line adaptation of Hidden Markov Models using incremental estimation models. In *European Conference on Speech Communication and Technology*, pages 1859–1862, 1997.
- B. Efron. Defining the curvature of a statistical problem (with applications to second order efficiency). *Annals of Statistics*, 3(6):1189–1242, 1975.

- R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A*, 222:309–768, 1922.
- A. Gunawardana. *The Information Geometry of EM Variants for Speech and Image Processing*. PhD thesis, The Johns Hopkins University, 2001.
- I.-T. Hsiao, A. Rangarajan, P. Khurd, and G. Gindi. Fast, globally convergent, reconstruction in emission tomography using COSEM, an incremental EM algorithm. *IEEE Transactions on Medical Imaging*, 2004. Submitted.
- M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 105–162. MIT Press, 1999.
- E. L. Lehmann. Efficient likelihood estimators. *The American Statistician*, 34(4):233–235, November 1980.
- F. Liese and I. Vajda. *Convex Statistical Distances*. Teubner Verlagsgesellschaft, Leipzig, 1987.
- G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1997.
- I. Meilijson. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society, Series B*, 51(1):127–138, 1989.
- R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*. Kluwer Academic Press, 1998.
- R. Salakhutdinov, S. T. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. In *Conference on Uncertainty in Artificial Intelligence*, volume 19, 2003.
- B. Thiesson, C. Meek, and D. Heckerman. Accelerating EM for large databases. *Machine Learning*, pages 279–299, 2001.
- A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, 20, 1949.
- C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.
- W. I. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall, 1969.