MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1520 January 17, 1995
C.B.C.L. Paper No. 111

On Convergence Properties of the EM Algorithm for Gaussian Mixtures

Lei Xu and Michael I. Jordan

jordan@ai.mit.edu

Abstract

We build up the mathematical connection between the "Expectation-Maximization" (EM) algorithm and gradient-based approaches for maximum likelihood learning of finite Gaussian mixtures. We show that the EM step in parameter space is obtained from the gradient via a projection matrix P, and we provide an explicit expression for the matrix. We then analyze the convergence of EM in terms of special properties of P and provide new results analyzing the effect that P has on the likelihood surface. Based on these mathematical results, we present a comparative discussion of the advantages and disadvantages of EM and other algorithms for the learning of Gaussian mixture models.

This report describes research done at the Center for Biological and Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Center is provided in part by a grant from the National Science Foundation under contract ASC–9217041. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00000-00-A-0000. The authors were also supported by the HK RGC Earmarked Grant CUHK250/94E, by a grant from the McDonnell-Pew Foundation, by a grant from ATR Human Information Processing Research Laboratories, by a grant from Siemens Corporation, and by grant N00014-90-1-0777 from the Office of Naval Research. Michael I. Jordan is an NSF Presidential Young Investigator.

1 Introduction

The "Expectation-Maximization" (EM) algorithm is a general technique for maximum likelihood (ML) or maximum a posteriori (MAP) estimation. The recent emphasis in the neural network literature on probabilistic models has led to increased interest in EM as a possible alternative to gradient-based methods for optimization. EM has been used for variations on the traditional theme of Gaussian mixture modeling (Ghahramani & Jordan, 1994; Nowlan, 1991; Xu & Jordan, 1993a, b; Tresp, Ahmad & Neuneier, 1994; Xu, Jordan & Hinton, 1994) and has also been used for novel chain-structured and tree-structured architectures (Bengio & Frasconi, 1995; Jordan & Jacobs, 1994). The empirical results reported in these papers suggest that EM has considerable promise as an optimization method for such architectures. Moreover, new theoretical results have been obtained that link EM to other topics in learning theory (Amari, 1994; Jordan & Xu, 1993; Neal & Hinton, 1993; Xu & Jordan, 1993c; Yuille, Stolorz & Utans, 1994).

Despite these developments, there are grounds for caution about the promise of the EM algorithm. One reason for caution comes from consideration of theoretical convergence rates, which show that EM is a first order algorithm.[1] More precisely, there are two key results available in the statistical literature on the convergence of EM. First, it has been established that under mild conditions EM is guaranteed to converge to a local maximum of the log likelihood l (Boyles, 1983; Dempster, Laird & Rubin, 1977; Redner & Walker, 1984; Wu, 1983). (Indeed the convergence is monotonic: $l(\Theta^{(k+1)}) \geq l(\Theta^{(k)})$, where $\Theta^{(k)}$ is the value of the parameter vector Θ at iteration k.) Second, considering EM as a mapping $\Theta^{(k+1)} = M(\Theta^{(k)})$ with fixed point $\Theta^* = M(\Theta^*)$, we have $\Theta^{(k+1)} - \Theta^* \approx \frac{\partial M(\Theta^*)}{\partial \Theta^*}(\Theta^{(k)} - \Theta^*)$ when $\Theta^{(k+1)}$ is near Θ^*, and thus

$$\|\Theta^{(k+1)} - \Theta^*\| \leq \|\frac{\partial M(\Theta^*)}{\partial \Theta^*}\| \cdot \|\Theta^{(k)} - \Theta^*\|,$$

with

$$\|\frac{\partial M(\Theta^*)}{\partial \Theta^*}\| \neq 0$$

almost surely. That is, EM is a first order algorithm.

The first-order convergence of EM has been cited in the statistical literature as a major drawback. Redner and Walker (1984), in a widely-cited article, argued that superlinear (quasi-Newton, method of scoring) and second-order (Newton) methods should generally be preferred to EM. They reported empirical results demonstrating the slow convergence of EM on a Gaussian mixture model problem for which the mixture components were not well separated. These results did not include tests of competing algorithms, however. Moreover, even though the convergence toward the "optimal" parameter values was slow in these experiments, the convergence in likelihood was rapid. Indeed, Redner and Walker acknowledge that their results show that "... even when the component populations in a mixture are poorly separated, the EM algorithm can be expected to produce in a very small number of iterations parameter values such that the mixture density determined by them reflects the sample data very well." In the context of the current literature on learning, in which the predictive aspect of data modeling is emphasized at the expense of the traditional Fisherian statistician's concern over the "true" values of parameters, such rapid convergence in likelihood is a major desideratum of a learning algorithm and undercuts the critique of EM as a "slow" algorithm.

In the current paper, we provide a comparative analysis of EM and other optimization methods. We emphasize the comparison between EM and other first-order methods (gradient ascent, conjugate gradient methods), because these have tended to be the methods of choice in the neural network literature. However, we also compare EM to superlinear and second-order methods. We argue that EM has a number of advantages, including its naturalness at handling the probabilistic constraints of mixture problems and its guarantees of convergence. We also provide new results suggesting that under appropriate conditions EM may in fact approximate a superlinear method; this would explain some of the promising empirical results that have been obtained (Jordan & Jacobs, 1994), and would further temper the critique of EM offered by Redner and Walker. The analysis in the current paper focuses on unsupervised learning; for related results in the supervised learning domain see Jordan and Xu (in press).

The remainder of the paper is organized as follows. We first briefly review the EM algorithm for Gaussian mixtures. The second section establishes a connection between EM and the gradient of the log likelihood. We then present a comparative discussion of the advantages and disadvantages of various optimization algorithms in the Gaussian mixture setting. We then present empirical results suggesting that EM regularizes the condition number of the effective Hessian. The fourth section presents a theoretical analysis of this empirical finding. The final section presents our conclusions.

2 The EM algorithm for Gaussian mixtures

We study the following probabilistic model:

$$P(x|\Theta) = \sum_{j=1}^{K} \alpha_j P(x|m_j, \Sigma_j), \tag{1}$$

and

$$P(x|m_j, \Sigma_j) = \frac{1}{(2\pi)^{d/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-m_j)^T \Sigma_j^{-1}(x-m_j)}$$

where $\alpha_j \geq 0$ and $\sum_{j=1}^{K} \alpha_j = 1$ and d is the dimensionality of the vector x. The parameter vector Θ consists of the mixing proportions α_j, the mean vectors m_j, and the covariance matrices Σ_j.

Given K and given N independent, identically distributed samples $\{x^{(t)}\}_1^N$, we obtain the following log

[1] An iterative algorithm is said to have a local convergence rate of order $q \geq 1$ if $\|\Theta^{(k+1)} - \Theta^*\| / \|\Theta^{(k)} - \Theta^*\|^q \leq r + o(\|\Theta^{(k)} - \Theta^*\|)$ for k sufficiently large.

likelihood:[2]

$$l(\Theta) = \log \prod_{t=1}^{N} P(x^{(t)}|\Theta) = \sum_{t=1}^{N} \log P(x^{(t)}|\Theta), \qquad (2)$$

which can be optimized via the following iterative algorithm (see, e.g, Dempster, Laird & Rubin, 1977):

$$\begin{aligned} \alpha_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)}{N} \qquad (3)\\ m_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t) x^{(t)}}{\sum_{t=1}^N h_j^{(k)}(t)} \\ \Sigma_j^{(k+1)} &= \frac{\sum_{t=1}^N h_j^{(k)}(t)[x^{(t)} - m_j^{(k+1)}][x^{(t)} - m_j^{(k+1)}]^T}{\sum_{t=1}^N h_j^{(k)}(t) x^{(t)}} \end{aligned}$$

where the posterior probabilities $h_j^{(k)}$ are defined as follows:

$$h_j^{(k)}(t) = \frac{\alpha_j^{(k)} P(x^{(t)}|m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}|m_i^{(k)}, \Sigma_i^{(k)})}.$$

3 Connection between EM and gradient ascent

In the following theorem we establish a relationship between the gradient of the log likelihood and the step in parameter space taken by the EM algorithm. In particular we show that the EM step can be obtained by premultiplying the gradient by a positive definite matrix. We provide an explicit expression for the matrix.

Theorem 1 *At each iteration of the EM algorithm Eq. (3), we have*

$$\begin{aligned} \mathcal{A}^{(k+1)} - \mathcal{A}^{(k)} &= P_{\mathcal{A}}^{(k)} \frac{\partial l}{\partial \mathcal{A}}|_{\mathcal{A}=\mathcal{A}^{(k)}} \qquad (4)\\ m_j^{(k+1)} - m_j^{(k)} &= P_{m_j}^{(k)} \frac{\partial l}{\partial m_j}|_{m_j = m_j^{(k)}} \qquad (5)\\ \mathrm{vec}[\Sigma_j^{(k+1)}] - \mathrm{vec}[\Sigma_j^{(k)}] &= P_{\Sigma_j}^{(k)} \frac{\partial l}{\partial \mathrm{vec}[\Sigma_j]}|_{\Sigma_j = \Sigma_j^{(k)}} \qquad (6) \end{aligned}$$

where

$$\begin{aligned} P_{\mathcal{A}}^{(k)} &= \frac{1}{N}\{\mathrm{diag}[\alpha_1^{(k)}, \cdots, \alpha_K^{(k)}] - \mathcal{A}^{(k)}(\mathcal{A}^{(k)})^T\} \qquad (7)\\ P_{m_j}^{(k)} &= \frac{\Sigma_j^{(k)}}{\sum_{t=1}^N h_j^{(k)}(t)} \qquad (8)\\ P_{\Sigma_j}^{(k)} &= \frac{2}{\sum_{t=1}^N h_j^{(k)}(t)} \Sigma_j^{(k)} \otimes \Sigma_j^{(k)} \qquad (9) \end{aligned}$$

where $\mathcal{A}$ denotes the vector of mixing proportions $[\alpha_1, \cdots, \alpha_K]^T$, j indexes the mixture components ($j = 1, \cdots, K$), k denotes the iteration number, "vec[B]" denotes the vector obtained by stacking the column vectors of the matrix B, and "$\otimes$" denotes the Kronecker product. Moreover, given the constraints $\sum_{j=1}^K \alpha_j^{(k)} = 1$ and $\alpha_j^{(k)} \geq 0$, $P_{\mathcal{A}}^{(k)}$ is a positive definite matrix and the matrices $P_{m_j}^{(k)}$ and $P_{\Sigma_j}^{(k)}$ are positive definite with probability one for N sufficiently large.

Proof. (1) We begin by considering the EM update for the mixing proportions α_i. From Eqs. (1) and (2), we have

$$\frac{\partial l}{\partial \mathcal{A}}|_{\mathcal{A}=\mathcal{A}^{(k)}} = \sum_{t=1}^N \frac{[P(x^{(t)}, \theta_1^{(k)}), \cdots, P(x^{(t)}, \theta_K^{(k)})]^T}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}, \theta_i^{(k)})}.$$

Premultiplying by $P_{\mathcal{A}}^{(k)}$, we obtain

$$\begin{aligned} &P_{\mathcal{A}}^{(k)} \frac{\partial l}{\partial \mathcal{A}}|_{\mathcal{A}=\mathcal{A}^{(k)}} \\ &= \frac{1}{N}\sum_{t=1}^N \frac{\{[\alpha_1^{(k)} P(x^{(t)}, \theta_1^{(k)}), \cdots]^T - \mathcal{A}^{(k)} \sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}, \theta_i^{(k)})\}}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}, \theta_i^{(k)})} \\ &= \frac{1}{N}\sum_{t=1}^N [h_1^{(k)}(t), \cdots, h_K^{(k)}(t)]^T - \mathcal{A}^{(k)}. \end{aligned}$$

The update formula for $\mathcal{A}$ in Eq. (3) can be rewritten as

$$\mathcal{A}^{(k+1)} = \mathcal{A}^{(k)} + \frac{1}{N}\sum_{t=1}^N [h_1^{(k)}(t), \cdots, h_K^{(k)}(t)]^T - \mathcal{A}^{(k)}.$$

Combining the last two equations establishes the update rule for $\mathcal{A}$ (Eq. 4). Furthermore, for an arbitrary vector u, we have $N u^T P_{\mathcal{A}}^{(k)} u = u^T \mathrm{diag}[\alpha_1^{(k)}, \cdots, \alpha_K^{(k)}]u - (u^T \mathcal{A}^{(k)})^2$. By Jensen's inequality we have

$$\begin{aligned} u^T \mathrm{diag}[\alpha_1^{(k)}, \cdots, \alpha_K^{(k)}]u &= \sum_{j=1}^K \alpha_j^{(k)} u_j^2 \\ &> (\sum_{j=1}^K \alpha_j^{(k)} u_j)^2 \\ &= (u^T \mathcal{A}^{(k)})^2. \end{aligned}$$

Thus, $u^T P_{\mathcal{A}}^{(k)} u > 0$ and $P_{\mathcal{A}}^{(k)}$ is positive definite given the constraints $\sum_{j=1}^K \alpha_j^{(k)} = 1$ and $\alpha_j^{(k)} \geq 0$ for all j.

(2) We now consider the EM update for the means m_i. It follows from Eqs. (1) and (2) that

$$\frac{\partial l}{\partial m_j}|_{m_j = m_j^{(k)}} = \sum_{t=1}^N h_j^{(k)}(t)(\Sigma_j^{(k)})^{-1}[x^{(t)} - m_j^{(k)}].$$

Premultiplying by $P_{m_j}^{(k)}$ yields

$$\begin{aligned} P_{m_j}^{(k)} \frac{\partial l}{\partial m_j}|_{m_j = m_j^{(k)}} &= \frac{1}{\sum_{t=1}^N h_j^{(k)}(t)} \sum_{t=1}^N h_j^{(k)}(t) x^{(t)} - m_j^{(k)} \\ &= m_j^{(k+1)} - m_j^{(k)}. \end{aligned}$$

From Eq. (3), we have $\sum_{t=1}^N h_j^{(k)}(t) > 0$; moreover, $\Sigma_j^{(k)}$ is positive definite with probability one assuming that N

[2]Although we focus on maximum likelihood (ML) estimation in this paper, it is straightforward to apply our results to maximum a posteriori (MAP) estimation by multiplying the likelihood by a prior.

is large enough such that the matrix is of full rank. Thus, it follows from Eq. (8) that $P_{m_j}^{(k)}$ is positive definite with probability one.

(3) Finally, we prove the third part of the theorem. It follows from Eqs. (1) and (2) that

$$\frac{\partial l}{\partial \Sigma_j}|_{\Sigma_j=\Sigma_j^{(k)}} = -\frac{1}{2}\sum_{t=1}^{N} h_j^{(k)}(t)(\Sigma_j^{(k)})^{-1}$$
$$\{\Sigma_j^{(k)} - [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)})]^T\}(\Sigma_j^{(k)})^{-1}.$$

With this in mind, we rewrite the EM update formula for $\Sigma_j^{(k)}$ as

$$\begin{aligned}\Sigma_j^{(k+1)} &= \Sigma_j^{(k)} + \frac{1}{\sum_{t=1}^{N} h_j^{(k)}(t)} \sum_{t=1}^{N} h_j^{(k)}(t)[x^{(t)} - m_j^{(k)}] \\ &\quad [x^{(t)} - m_j^{(k)}]^T - \Sigma_j^{(k)} \\ &= \Sigma_j^{(k)} + \frac{2\Sigma_j^{(k)}}{\sum_{t=1}^{N} h_j^{(k)}(t)} V_{\Sigma_j}\Sigma_j^{(k)},\end{aligned}$$

where

$$\begin{aligned}V_{\Sigma_j} &= -\frac{1}{2}\sum_{t=1}^{N} h_j^{(k)}(t)(\Sigma_j^{(k)})^{-1} \\ &\quad \{\Sigma_j^{(k)} - [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T\}(\Sigma_j^{(k)})^{-1} \\ &= \frac{\partial l}{\partial \Sigma_j}|_{\Sigma_j=\Sigma_j^{(k)}}.\end{aligned}$$

That is, we have

$$\Sigma_j^{(k+1)} = \Sigma_j^{(k)} + \frac{2\Sigma_j^{(k)}}{\sum_{t=1}^{N} h_j^{(k)}(t)} \frac{\partial l}{\partial \Sigma_j}|_{\Sigma_j=\Sigma_j^{(k)}} \Sigma_j^{(k)}.$$

Utilizing the identity $\mathrm{vec}[ABC] = (C^T \otimes A)\mathrm{vec}[B]$, we obtain

$$\mathrm{vec}[\Sigma_j^{(k+1)}] = \mathrm{vec}[\Sigma_j^{(k)}]$$
$$+ \frac{2}{\sum_{t=1}^{N} h_j^{(k)}(t)} (\Sigma_j^{(k)} \otimes \Sigma_j^{(k)}) \frac{\partial l}{\partial \Sigma_j}|_{\Sigma_j=\Sigma_j^{(k)}}.$$

Thus $P_{\Sigma_j}^{(k)} = \frac{2}{\sum_{t=1}^{N} h_j^{(k)}(t)}(\Sigma_j^{(k)} \otimes \Sigma_j^{(k)})$. Moreover, for an arbitrary matrix U, we have

$$\begin{aligned}&\mathrm{vec}[U]^T(\Sigma_j^{(k)} \otimes \Sigma_j^{(k)})\mathrm{vec}[U] \\ &= \mathrm{tr}(\Sigma_j^{(k)} U \Sigma_j^{(k)} U^T) \\ &= \mathrm{tr}((\Sigma_j^{(k)} U)^T (\Sigma_j^{(k)} U)) \\ &= \mathrm{vec}[\Sigma_j^{(k)} U]^T \mathrm{vec}[\Sigma_j^{(k)} U] \\ &\geq 0,\end{aligned}$$

where equality holds only when $\Sigma_j^{(k)} U = 0$ for all U. Equality is impossible, however, since $\Sigma_j^{(k)}$ is positive definite with probability one N is sufficiently large. Thus it follows from Eq. (9) and $\sum_{t=1}^{N} h_j^{(k)}(t) > 0$ that $P_{\Sigma_j}^{(k)}$ is positive definite with probability one. □

Using the notation

$$\Theta = [m_1^T, \cdots, m_K^T, \mathrm{vec}[\Sigma_1]^T, \cdots, \mathrm{vec}[\Sigma_K]^T, \mathcal{A}^T]^T,$$

and $P(\Theta) = \mathrm{diag}[P_{m_1}, \cdots, P_{m_K}, P_{\Sigma_1}, \cdots, P_{\Sigma_K}, P_{\mathcal{A}}]$, we can combine the three updates in Theorem 1 into a single equation:

$$\Theta^{(k+1)} = \Theta^{(k)} + P(\Theta^{(k)})\frac{\partial l}{\partial \Theta}|_{\Theta=\Theta^{(k)}}, \qquad (10)$$

Under the conditions of Theorem 1, $P(\Theta^{(k)})$ is a positive definite matrix with probability one. Recalling that for a positive definite matrix B, we have $\frac{\partial l}{\partial \Theta}^T B \frac{\partial l}{\partial \Theta} > 0$, we have the following corollary:

Corollary 1 *For each iteration of the EM algorithm given by Eq.(3), the search direction $\Theta^{(k+1)} - \Theta^{(k)}$ has a positive projection on the gradient of l.*

That is, the EM algorithm can be viewed as a variable metric gradient ascent algorithm for which the projection matrix $P(\Theta^{(k)})$ changes at each iteration as a function of the current parameter value $\Theta^{(k)}$.

Our results extend earlier results due to Baum and Sell (1968). Baum and Sell studied recursive equations of the following form:

$$\begin{aligned}x^{(k+1)} &= T(x^{(k)}) \\ T(x^{(k)}) &= [T(x^{(k)})_1, \cdots, T(x^{(k)})_K] \\ T(x^{(k)}))_i &= \frac{x_i^{(k)}\partial J/\partial x_i^{(k)}}{\sum_{i=1}^{K} x_i^{(k)}\partial J/\partial x_i^{(k)}}\end{aligned}$$

where $x_i^{(k)} \geq 0$, $\sum_{i=1}^{K} x_i^{(k)} = 1$, where J is a polynomial in $x_i^{(k)}$ having positive coefficients. They showed that the search direction of this recursive formula, i.e., $T(x^{(k)}) - x^{(k)}$, has a positive projection on the gradient of of J with respect to the $x^{(k)}$ (see also Levinson, Rabiner & Sondhi, 1983). It can be shown that Baum and Sell's recursive formula implies the EM update formula for $\mathcal{A}$ in a Gaussian mixture. Thus, the first statement in Theorem 1 is a special case of Baum and Sell's earlier work. However, Baum and Sell's theorem is an existence theorem and does not provide an explicit expression for the matrix $P_{\mathcal{A}}$ that transforms the gradient direction into the EM direction. Our theorem provides such an explicit form for $P_{\mathcal{A}}$. Moreover, we generalize Baum and Sell's results to handle the updates for m_j and Σ_j, and we provide explicit expressions for the positive definite transformation matrices P_{m_j} and P_{Σ_j} as well.

It is also worthwhile to compare the EM algorithm to other gradient-based optimization methods. *Newton's method* is obtained by premultiplying the gradient by the inverse of the Hessian of the log likelihood:

$$\Theta^{(k+1)} = \Theta^{(k)} + H(\Theta^{(k)})^{-1}\frac{\partial l}{\partial \Theta^{(k)}}. \qquad (11)$$

Newton's method is the method of choice when it can be applied, but the algorithm is often difficult to use in practice. In particular, the algorithm can diverge when the Hessian becomes nearly singular; moreover, the computational costs of computing the inverse Hessian at each step can be considerable. An alternative is to approximate the inverse by a recursively updated matrix $B^{(k+1)} = B^{(k)} + \eta \Delta B^{(k)}$. Such a modification is called a *quasi-Newton method.* Conventional quasi-Newton methods are unconstrained optimization methods, however, and must be modified in order to be used in the mixture setting (where there are probabilistic constraints on the parameters). In addition, quasi-Newton methods generally require that a one-dimensional search be performed at each iteration in order to guarantee convergence. The EM algorithm can be viewed as a special form of quasi-Newton method in which the projection matrix $P(\Theta^{(k)})$ in Eq. (10) plays the role of $B^{(k)}$. As we discuss in the remainder of the paper, this particular matrix has a number of favorable properties that make EM particularly attractive for optimization in the mixture setting.

4 Constrained optimization and general convergence

An important property of the matrix P is that the EM step in parameter space automatically satisfies the probabilistic constraints of the mixture model in Eq. (1). The domain of Θ contains two regions that embody the probabilistic constraints: $\mathcal{D}_1 = \{\Theta : \sum_{j=1}^K \alpha_j^{(k)} = 1\}$ and $\mathcal{D}_2 = \{\Theta : \alpha_j^{(k)} \geq 0,\ \Sigma_j$ positive definite$\}$. For the EM algorithm the update for the mixing proportions α_j can be rewritten as follows:

$$\alpha_j^{(k+1)} = \frac{1}{N} \sum_{t=1}^N \frac{\alpha_j^{(k)} P(x^{(t)}|m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}|m_i^{(k)}, \Sigma_i^{(k)})}.$$

It is obvious that the iteration stays within $\mathcal{D}_1$. Similarly, the update for Σ_j can be rewritten as:

$$\begin{aligned}\Sigma_j^{(k+1)} &= \frac{1}{\sum_{t=1}^N h_j^{(k)}(t)} \sum_{t=1}^N \frac{\alpha_j^{(k)} P(x^{(t)}|m_j^{(k)}, \Sigma_j^{(k)})}{\sum_{i=1}^K \alpha_i^{(k)} P(x^{(t)}|m_i^{(k)}, \Sigma_i^{(k)})} \\ &\quad [x^{(t)} - m_j^{(k)}][x^{(t)} - m_j^{(k)}]^T\end{aligned}$$

which stays within $\mathcal{D}_2$ for N sufficiently large.

Whereas EM automatically satisfies the probabilistic constraints of a mixture model, other optimization techniques generally require modification to satisfy the constraints. One approach is to modify each iterative step to keep the parameters within the constrained domain. A number of such techniques have been developed, including feasible direction methods, active sets, gradient projection, reduced-gradient, and linearly-constrained quasi-Newton. These constrained methods all incur extra computational costs to check and maintain the constraints and, moreover, the theoretical convergence rates for such constrained algorithms need not be the same as that for the corresponding unconstrained algorithms. A second approach is to transform the constrained optimization problem into an unconstrained problem before using the unconstrained method. This can be accomplished via penalty and barrier functions, Lagrangian terms, or re-parameterization. Once again, the extra algorithmic machinery renders simple comparisons based on unconstrained convergence rates problematic. Moreover, it is not easy to meet the constraints on the covariance matrices in the mixture using such techniques.

A second appealing property of $P(\Theta^{(k)})$ is that each iteration of EM is guaranteed to increase the likelihood (i.e., $l(\Theta^{(k+1)}) \geq l(\Theta^{(k)})$). This monotonic convergence of the likelihood is achieved without step-size parameters or line searches. Other gradient-based optimization techniques, including gradient descent, quasi-Newton, and Newton's method, do not provide such a simple theoretical guarantee, even assuming that the constrained problem has been transformed into an unconstrained one. For gradient ascent, the step size η must be chosen to ensure that $\|\Theta^{(k+1)} - \Theta^{(k-1)}\| / \|(\Theta^{(k)} - \Theta^{(k-1)})\| \leq \|I + \eta H(\Theta^{(k-1)}))\| < 1$. This requires a one-dimensional line search or an optimization of η at each iteration, which requires extra computation which can slow down the convergence. An alternative is to fix η to a very small value which generally makes $\|I + \eta H(\Theta^{(k-1)}))\|$ close to one and results in slow convergence. For Newton's method, the iterative process is usually required to be near a solution, otherwise the Hessian may be indefinite and the iteration may not converge. Levenberg-Marquardt methods handle the indefinite Hessian matrix problem; however, a one-dimensional optimization or other form of search is required for a suitable scalar to be added to the diagonal elements of Hessian. Fisher scoring methods can also handle the indefinite Hessian matrix problem, but for non-quadratic nonlinear optimization Fisher scoring requires a stepsize η that obeys $\|I + \eta BH(\Theta^{(k-1)}))\| < 1$, where B is the Fisher information matrix. Thus, problems similar to those of gradient ascent arise here as well. Finally, for the quasi-Newton methods or conjugate gradient methods, a one-dimensional line search is required at each iteration. In summary, all of these gradient-based methods incur extra computational costs at each iteration, rendering simple comparisons based on local convergence rates unreliable.

For large scale problems, algorithms that change the parameters immediately after each data point ("on-line algorithms") are often significantly faster in practice than batch algorithms. The popularity of gradient descent algorithms for neural networks is in part to the ease of obtaining on-line variants of gradient descent. It is worth noting that on-line variants of the EM algorithm can be derived (Neal & Hinton, 1993, Titterington, 1984), and this is a further factor that weighs in favor of EM as compared to conjugate gradient and Newton methods.

5 Convergence rate comparisons

In this section, we provide a comparative theoretical discussion of the convergence rates of constrained gradient ascent and EM.

For gradient ascent a local convergence result can by obtained by Taylor expanding the log likelihood around the maximum likelihood estimate Θ^*. For sufficiently large k we have:

$$\|\Theta^{(k+1)} - \Theta^*\| \leq \|I + \eta H(\Theta^*))\| \|(\Theta^{(k)} - \Theta^*)\| \quad (12)$$

and

$$\|I+\eta H(\Theta^*)\| \le \lambda_M[I+\eta H(\Theta^*)] = r, \qquad (13)$$

where H is the Hessian of l, η is the step size, and $r = \max\{|1-\eta\lambda_M[-H(\Theta^*)]|, \; |1-\eta\lambda_m[-H(\Theta^*)]|\}$, where $\lambda_M[A]$ and $\lambda_m[A]$ denote the largest and smallest eigenvalues of A, respectively.

Smaller values of r correspond to faster convergence rates. To guarantee convergence, we require $r<1$ or $0<\eta<2/\lambda_M[-H(\Theta^*)]$. The minimum possible value of r is obtained when $\eta = 1/\lambda_M[H(\Theta^*)]$ with

$$\begin{aligned} r_{min} &= 1-\lambda_m[H(\Theta^*)]/\lambda_M[H(\Theta^*)] \\ &\equiv 1-\kappa^{-1}[H(\Theta^*)], \end{aligned}$$

where $\kappa[H] = \lambda_M[H]/\lambda_m[H]$ is the *condition number* of H. Larger values of the condition number correspond to slower convergence. When $\kappa[H]=1$ we have $r_{min}=0$, which corresponds to a superlinear rate of convergence. Indeed, Newton's method can be viewed as a method for obtaining a more desirable condition number—the inverse Hessian H^{-1} balances the Hessian H such that the resulting condition number is one. Effectively, Newton can be regarded as gradient ascent on a new function with an effective Hessian that is the identity matrix: $H_{eff} = H^{-1}H = I$. In practice, however, $\kappa[H]$ is usually quite large. The larger $\kappa[H]$ is, the more difficult it is to compute H^{-1} accurately. Hence it is difficult to balance the Hessian as desired. In addition, as we mentioned in the previous section, the Hessian varies from point to point in the parameter space, and at each iteration we need recompute the inverse Hessian. Quasi-Newton methods approximate $H(\Theta^{(k)})^{-1}$ by a positive matrix $B^{(k)}$ that is easy to compute.

The discussion thus far has treated unconstrained optimization. In order to compare gradient ascent with the EM algorithm on the constrained mixture estimation problem, we consider a gradient projection method:

$$\Theta^{(k+1)} = \Theta^{(k)} + \eta\Pi_k \frac{\partial l}{\partial \Theta^{(k)}} \qquad (14)$$

where Π_k is the projection matrix that projects the gradient $\frac{\partial l}{\partial\Theta^{(k)}}$ into $\mathcal{D}_1$. This gradient projection iteration will remain in $\mathcal{D}_1$ as long as the initial parameter vector is in $\mathcal{D}_1$. To keep the iteration within $\mathcal{D}_2$, we choose an initial $\Theta^{(0)} \in \mathcal{D}_2$ and keep η sufficiently small at each iteration.

Suppose that $E = [e_1, \cdots, e_m]$ are a set of independent unit basis vectors that span the space $\mathcal{D}_1$. In this basis, $\Theta^{(k)}$ and $\Pi_k\frac{\partial l}{\partial\Theta^{(k)}}$ become $\Theta_c^{(k)} = E^T\Theta^{(k)}$ and $\frac{\partial l}{\partial\Theta^{(k)}}_c = E^T\frac{\partial l}{\partial\Theta^{(k)}}$, respectively, with $\|\Theta_c^{(k)} - \Theta_c^*\| = \|\Theta^{(k)} - \Theta^*\|$. In this representation the projective gradient algorithm Eq. (14) becomes simple gradient ascent: $\Theta_c^{(k+1)} = \Theta_c^{(k)} + \eta\frac{\partial l}{\partial\Theta_c^{(k)}}$. Moreover, Eq. (12) becomes $\|\Theta^{(k+1)} - \Theta^*\| \le \|E^T(I+\eta H(\Theta^*))\| \|(\Theta^{(k)} - \Theta^*)\|$. As a result, the convergence rate is bounded by

$$\begin{aligned} r_c &= \|E^T(I+\eta H(\Theta^*))\| \\ &\le \sqrt{\lambda_M[E^T(I+\eta H(\Theta^*))(I+\eta H(\Theta^*))^T E]} \\ &= \sqrt{\lambda_M[E^T(I+2\eta H(\Theta^*)+\eta^2 H^2(\Theta^*))E]}. \end{aligned}$$

Since $H(\Theta^*)$ is negative definite, we obtain

$$r_c \le \sqrt{1+\eta^2\lambda_M^2[-H_c] - 2\eta\lambda_m[-H_c]}. \qquad (15)$$

In this equation $H_c = E^T H(\Theta)E$ is the Hessian of l restricted to $\mathcal{D}_1$.

We see from this derivation that the convergence speed depends on $\kappa[H_c] = \lambda_M[-H_c]/\lambda_m[-H_c]$. When $\kappa[H_c]=1$, we have $\sqrt{1+\eta^2\lambda_M^2(-H_c) - 2\eta\lambda_m[-H_c]} = 1-\eta\lambda[-H_c]$, which in principle can be made to equal zero if η is selected appropriately. In this case, a superlinear rate is obtained. Generally, however, $\kappa[H_c] \neq 1$, with smaller values of $\kappa[H_c]$ corresponding to faster convergence.

We now turn to an analysis of the EM algorithm. As we have seen EM keeps the parameter vector within $\mathcal{D}_1$ automatically. Thus, in the new basis the connection between EM and gradient ascent (cf. Eq. (10)) becomes

$$\Theta_c^{(k+1)} = \Theta_c^{(k)} + E^T P(\Theta^{(k)})\frac{\partial l}{\partial\Theta}$$

and we have

$$\|\Theta^{(k+1)} - \Theta^*\| \le \|E^T(I+PH(\Theta^*))\| \|(\Theta^{(k)} - \Theta^*)\|$$

with

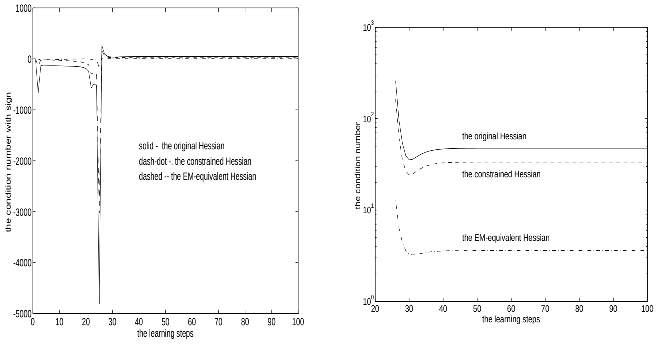
$$\begin{aligned} r_c &= \|E^T(I+PH(\Theta^*))\| \\ &\le \sqrt{\lambda_M[E^T(I+PH(\Theta^*))(I+PH(\Theta^*))^T E]}. \end{aligned}$$

The latter equation can be further manipulated to yield:

$$r_c \le \sqrt{1+\lambda_M^2[E^TPHE] - 2\lambda_m[-E^TPHE]}. \qquad (16)$$

Thus we see that the convergence speed of EM depends on $\kappa[E^TPHE] = \lambda_M[E^TPHE]/\lambda_m[E^TPHE]$. When $\kappa[E^TPHE] = 1$, $\lambda_M[E^TPHE] = 1$, we have $\sqrt{1+\lambda_M^2[E^TPHE] - 2\lambda_m[-E^TPHE]} = (1-\lambda_M[-E^TPHE]) = 0$. In this case, a superlinear rate is obtained. We discuss the possibility of obtaining superlinear convergence with EM in more detail below.

These results show that the convergence of gradient ascent and EM both depend on the shape of the log likelihood as measured by the condition number. When $\kappa[H]$ is near one, the configuration is quite regular, and the update direction points directly to the solution yielding fast convergence. When $\kappa[H]$ is very large, the l surface has an elongated shape, and the search along the update direction is a zigzag path, making convergence very slow. The key idea of Newton and quasi-Newton methods is to reshape the surface. The nearer it is to a ball shape (Newton's method achieves this shape in the ideal case), the better the convergence. Quasi-Newton methods aim to achieve an effective Hessian whose condition number is as close as possible to one. Interestingly, the results that we now present suggest that the projection matrix P for the EM algorithm also serves to effectively reshape the likelihood yielding an effective condition number that tends to one. We first present empirical results that support this suggestion and then present a theoretical analysis.



(a) (b)

Figure 1: Experimental results for the estimation of the parameters of a two-component Gaussian mixture. (a) The condition numbers as a function of the iteration number. (b) A zoomed version of (a) after discarding the first 25 iterations. The terminology 'original, constrained, and EM-equivalent Hessians' refers to the matrices H, E^THE, and E^TPHE respectively.

We sampled 1000 points from a simple finite mixture model given by

$$p(x) = \alpha_1 p_1(x) + \alpha_2 p_2(x)$$

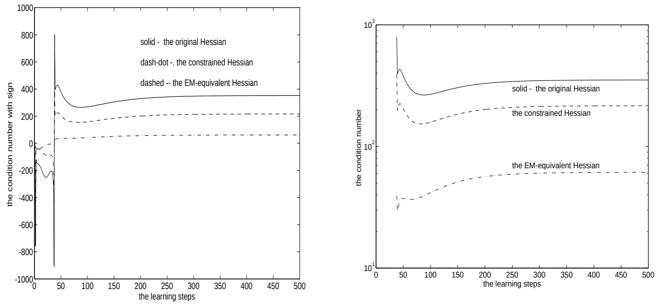
where

$$p_i(x) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\{-\frac{1}{2}\frac{(x-m_i)^2}{\sigma_i^2}\}.$$

The parameter values were as follows: $\alpha_1 = 0.7170$, $\alpha_2 = 0.2830$, $m_1 = -2$, $m_2 = 2$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$. We ran both the EM algorithm and gradient ascent on the data. At each step of the simulation, we calculated the condition number of the Hessian ($\kappa[H(\Theta^{(k)})]$), the condition number determining the rate of convergence of the gradient algorithm ($\kappa[E^TH(\Theta^{(k)})E]$), and the condition number determining the rate of convergence of EM ($\kappa[E^TP(\Theta^{(k)})H(\Theta^{(k)})E]$). We also calculated the largest eigenvalues of the matrices $H(\Theta^{(k)})$, $E^TH(\Theta^{(k)})E$, and $E^TP(\Theta^{(k)})H(\Theta^{(k)})E$. The results are shown in Fig. 1. As can be seen in Fig. 1(a), the condition numbers change rapidly in the vicinity of the 25th iteration and the corresponding Hessian matrices become indefinite. Afterward, the Hessians quickly become definite and the condition numbers converge.[3] As shown in Fig. 1(b), the condition numbers converge toward the values $\kappa[H(\Theta^{(k)})] = 47.5$, $\kappa[E^TH(\Theta^{(k)})E] = 33.5$, and $\kappa[E^TP(\Theta^{(k)})H(\Theta^{(k)})E] = 3.6$. That is, the matrix P has greatly reduced the condition number, by factors of 9 and 15. This significantly improves the shape of l and speeds up the convergence.

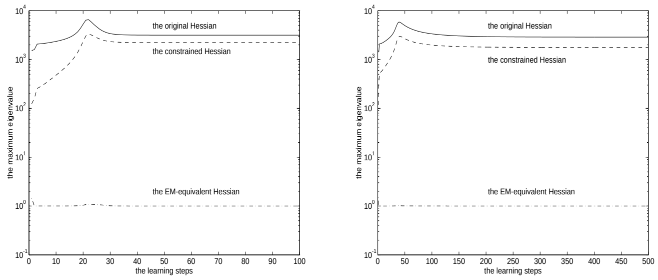
We ran a second experiment in which the means of the component Gaussians were $m_1 = -1$ and $m_2 = 1$. The results are similar to those shown in Fig. 1. Since the distance between two distributions is reduced into half,

[3]Interestingly, the EM algorithm converges soon afterward as well, showing that for this problem EM spends little time in the region of parameter space in which a local analysis is valid.



(a) (b)

Figure 2: Experimental results for the estimation of the parameters of a two-component Gaussian mixture (cf. Fig. 1). The separation of the Gaussians is half the separation in Fig. 1.



(a) (b)

Figure 3: The largest eigenvalues of the matrices H, E^THE, and E^TPHE plotted as a function of the number of iterations. The plot in (a) is for the experiment in Fig. 1; (b) is for the experiment reported in Fig. 2.

the shape of l becomes more irregular. The condition number $\kappa[H(\Theta^{(k)})]$ increases to 352, $\kappa[E^TH(\Theta^{(k)})E]$ increases to 216, and $\kappa[E^TP(\Theta^{(k)})H(\Theta^{(k)})E]$ increases to 61. We see once again a significant improvement in the case of EM, by factors of 4 and 6.

Fig. 3 shows that the matrix P has also reduced the largest eigenvalues of the Hessian from between 2000 to 3000 to around 1. This demonstrates clearly the stable convergence that is obtained via EM, without a line search or the need for external selection of a learning stepsize.

In the remainder of the paper we provide some theoretical analyses that attempt to shed some light on these empirical results. To illustrate the issues involved, consider a degenerate mixture problem in which the mixture has a single component. (In this case $\alpha_1 = 1$.) Let us furthermore assume that the covariance matrix is fixed (i.e., only the mean vector m is to be estimated). The Hessian with respect to the mean m is $H = -N\Sigma^{-1}$ and the EM projection matrix P is Σ/N. For gradient ascent, we have $\kappa[E^THE] = \kappa[\Sigma^{-1}]$, which is larger than one whenever $\Sigma \neq cI$. EM, on the other hand, achieves a condition number of one exactly ($\kappa[E^TPHE] = \kappa[PH] = \kappa[I] = 1$ and $\lambda_M[E^TPHE] = 1$). Thus, EM and Newton's method are the same for this simple quadratic problem. For general non-quadratic optimization problems, Newton retains the quadratic assumption, yielding fast convergence but possible divergence. EM is a more conservative algorithm that retains the convergence guarantee but also maintains quasi-Newton behavior. We now analyze this behavior in more detail. We consider the special case of estimating the means in a Gaussian mixture when the Gaussians are well separated.

Theorem 2 *Consider the EM algorithm in Eq. (3), where the parameters α_j and Σ_j are assumed to be known. Assume that the K Gaussian distributions are well separated, such that for sufficiently large k the posterior probabilities $h_j^{(k)}(t)$ are nearly zero or one. For such k, the condition number associated with EM is always smaller than the condition number associated with gradient ascent. That is:*

$$\kappa[E^T P(\Theta^{(k)})H(\Theta^{(k)})E] < \kappa[E^T H(\Theta^{(k)})E].$$

Furthermore, $\lambda_M[E^T P(\Theta^{(k)})H(\Theta^{(k)})E]$ approaches one as k goes to infinity.

Proof. The Hessian is

$$H = \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1K} \\ H_{21} & H_{22} & \cdots & H_{2K} \\ \vdots & \vdots & & \vdots \\ H_{K1} & H_{K2} & \cdots & H_{KK} \end{bmatrix} \qquad (17)$$

where

$$\begin{aligned} H_{ij} &\equiv \frac{\partial^2 l}{\partial m_i \partial m_j^T} \qquad (18) \\ &= -(\Sigma_j^{(k)})^{-1} \sum_{t=1}^{N} \delta_{ij} h_j^{(k)}(t) + (\Sigma_j^{(k)})^{-1} \\ &\quad [\sum_{t=1}^{N} \gamma_{ij}(x^{(t)})(x^{(t)} - m_j)(x^{(t)} - m_i)^T](\Sigma_i^{(k)})^{-1} \end{aligned}$$

with $\gamma_{ij}(x^{(t)}) = (\delta_{ij} - h_i^{(k)}(t))h_j^{(k)}(t)$. The projection matrix P is

$$P^{(k)} = \mathrm{diag}[P_{11}^{(k)}, \cdots, P_{KK}^{(k)}],$$

where

$$P_{jj}^{(k)} = \frac{\Sigma_j^{(k)}}{\sum_{t=1}^{N}} h_j^{(k)}(t).$$

Given that $h_j^{(k)}(t)(1 - h_j^{(k)}(t))$ is negligible for sufficiently large k, the second term in Eq. (19) can be neglected, yielding $H_{ii} = -(\Sigma_j^{(k)})^{-1} \sum_{t=1}^{N} h_j^{(k)}(t)$ and $H = \mathrm{diag}[H_{11}, \cdots, H_{KK}]$. This implies that $PH = -I$, and thus $\kappa[PH] = 1$, whereas $\kappa[H] \neq 1$. □

This theorem, although restrictive in its assumptions, gives some indication as to why the projection matrix in the EM algorithm appears to condition the Hessian, yielding improved convergence. In fact, we conjecture that the theorem can be extended to apply more widely, in particular to the case of the full EM update in which the mixing proportions and covariances are estimated, and also, within limits, to cases in which the means are not well separated. To obtain an initial indication as to possible conditions that can be usefully imposed on the separation of the mixture components, we have studied the case in which the second term in Eq. (19) is neglected only for H_{ii} and is retained for the H_{ij} components, where $j \neq i$. Consider, for example, the case of a univariate mixture having two mixture components. For fixed mixing proportions and fixed covariances, the Hessian matrix (Eq. 17) becomes:

$$H = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \end{bmatrix},$$

and the projection matrix (Eq. 19) becomes:

$$P = \begin{bmatrix} -h_{11}^{-1} & 0 \\ 0 & -h_{22}^{-1} \end{bmatrix},$$

where

$$h_{ii} = -\frac{1}{\sigma_i^{2(k)}} \sum_{t=1}^{N} h_i^{(k)}(t), \; i = 1, 2$$

and

$$h_{ij} = \frac{1}{\sigma_i^{2(k)} \sigma_j^{2(k)}} \sum_{t=1}^{N} (1 - h_i^{(k)}(t)) h_j^{(k)}(t)(x^{(t)} - m_j)^T (x^{(t)} - m_i),$$

for $i \neq j = 1, 2$. If H is negative definite, (i.e., $h_{11}h_{22} - h_{12}h_{21} < 0$), then we can show that the conclusions of Theorem 2 remain true, even for Gaussians that are not necessarily well-separated. The proof is achieved via the following lemma:

Lemma 1 *Consider the positive definite matrix*

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

For the diagonal matrix $B = \mathrm{diag}[\sigma_{11}^{-1}, \sigma_{22}^{-1}]$, we have $\kappa[B\Sigma] < \kappa[\Sigma]$.

Proof. The eigenvalues of Σ are the roots of $(\sigma_{11} - \lambda)(\sigma_{22} - \lambda) - \sigma_{21}\sigma_{12} = 0$, which gives

$$\begin{aligned} \lambda_M &= \frac{\sigma_{11} + \sigma_{22} + \gamma}{2} \\ \lambda_m &= \frac{\sigma_{11} + \sigma_{22} - \gamma}{2} \\ \gamma &= \sqrt{(\sigma_{11} + \sigma_{22})^2 - 4(\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12})} \end{aligned}$$

and

$$\kappa[\Sigma] = \frac{\sigma_{11} + \sigma_{22} + \gamma}{\sigma_{11} + \sigma_{22} - \gamma}$$

The condition number $\kappa[\Sigma]$ can be written as $\kappa[\Sigma] = (1 + s)/(1 - s) \equiv f(s)$, where s is defined as follows:

$$s = \sqrt{1 - \frac{4(\sigma_{11}\sigma_{22} - \sigma_{21}\sigma_{12})}{(\sigma_{11} + \sigma_{22})^2}}.$$

Furthermore, the eigenvalues of $B\Sigma$ are the roots of $(1-\lambda)(1-\lambda)-(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22}) = 0$, which gives $\lambda_M = 1+\sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$ and $\lambda_m = 1-\sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$. Thus, defining $r = \sqrt{(\sigma_{21}\sigma_{12})/(\sigma_{11}\sigma_{22})}$, we have $\kappa[B\Sigma] = (1+r)/(1-r) = f(r)$.

We now examine the quotient s/r:

$$\frac{s}{r} = \frac{1}{r}\sqrt{1-\frac{4(1-r^2)}{(\sigma_{11}+\sigma_{22})^2/(\sigma_{11}\sigma_{22})}}$$

Given that $(\sigma_{11}+\sigma_{22})^2/(\sigma_{11}\sigma_{22}) \geq 4$, we have $\frac{s}{r} > \frac{1}{r}\sqrt{1-(1-r^2)} = 1$. That is, $s > r$. Since $f(x) = (1+x)/(1-x)$ is a monotonically increasing function for $x > 0$, we have $f(s) > f(r)$. Therefore, $\kappa[B\Sigma] < \kappa[\Sigma]$. $\square$

We think that it should be possible to generalize this lemma beyond the univariate, two-component case, thereby weakening the conditions on separability in Theorem 2 in a more general setting.

6 Conclusions

In this paper we have provided a comparative analysis of algorithms for the learning of Gaussian mixtures. We have focused on the EM algorithm and have forged a link between EM and gradient methods via the projection matrix P. We have also analyzed the convergence of EM in terms of properties of the matrix P and the effect that P has on the likelihood surface.

EM has a number of properties that make it a particularly attractive algorithm for mixture models. It enjoys automatic satisfaction of probabilistic constraints, monotonic convergence without the need to set a learning rate, and low computational overhead. Although EM has the reputation of being a slow algorithm, we feel that in the mixture setting the slowness of EM has been overstated. Although EM can indeed converge slowly for problems in which the mixture components are not well separated, the Hessian is poorly conditioned for such problems and thus other gradient-based algorithms (including Newton's method) are also likely to perform poorly. Moreover, if one's concern is convergence in likelihood, then EM generally performs well even for these ill-conditioned problems. Indeed the algorithm provides a certain amount of safety in such cases, despite the poor conditioning. It is also important to emphasize that the case of poorly separated mixture components can be viewed as a problem in model selection (too many mixture components are being included in the model), and should be handled by regularization techniques.

The fact that EM is a first order algorithm certainly implies that EM is no panacea, but does not imply that EM has no advantages over gradient ascent or superlinear methods. First, it is important to appreciate that convergence rate results are generally obtained for unconstrained optimization, and are not necessarily indicative of performance on constrained optimization problems. Also, as we have demonstrated, there are conditions under which the condition number of the effective Hessian of the EM algorithm tends toward one, showing that EM can approximate a superlinear method. Finally, in cases of a poorly conditioned Hessian, superlinear convergence is not necessarily a virtue. In such cases many optimization schemes, including EM, essentially revert to gradient ascent.

We feel that EM will continue to play an important role in the development of learning systems that emphasize the predictive aspect of data modeling. EM has indeed played a critical role in the development of hidden Markov models (HMM's), an important example of predictive data modeling.[4] EM generally converges rapidly in this setting. Similarly, in the case of hierarchical mixtures of experts the empirical results on convergence in likelihood have been quite promising (Jordan & Jacobs, 1994; Waterhouse & Robinson, 1994). Finally, EM can play an important conceptual role as an organizing principle in the design of learning algorithms. Its role in this case is to focus attention on the "missing variables" in the problem. This clarifies the structure of the algorithm and invites comparisons with statistical physics, where missing variables often provide a powerful analytic tool.

7 References

Amari, S. (in press) Information geometry of the EM and em algorithms for neural networks, *Neural Networks*.

Baum, L.E., and Sell, G.R. (1968), Growth transformation for functions on manifolds, *Pac. J. Math., 27*, 211-227.

Bengio, Y., and Frasconi, P., (1995), An input-output HMM architecture. *Advances in Neural Information Processing Systems 6*, eds., Tesauro, G., Touretzky, D.S., and Alspector, J., San Mateo, CA: Morgan Kaufmann.

Boyles, R.A. (1983), On the convergence of the EM algorithm, *J. of Royal Statistical Society, B45*, No.1, 47-50.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. of Royal Statistical Society, B39*, 1-38.

Ghahramani, Z, and Jordan, M.I. (1994), Function approximation via density estimation using the EM approach, *Advances in Neural Information Processing Systems 6*, eds., Cowan, J.D., Tesauro, G., and Alspector, J., San Mateo, CA: Morgan Kaufmann, 120-127.

Jordan, M.I. and Jacobs, R.A. (1994), Hierarchical mixtures of experts and the EM algorithm. *Neural Computation 6*, 181-214.

Jordan, M.I. and Xu, L. (in press), Convergence results for the EM approach to mixtures-of-experts architectures, *Neural Networks*.

Levinson, S.E., Rabiner, L.R., and Sondhi, M.M. (1983), An introduction to the application of the theory of

[4] In most applications of HMM's, the "parameter estimation" process is employed solely to yield models with high likelihood; the parameters are not generally endowed with a particular meaning.

probabilistic functions of Markov process to automatic speech recognition, *The Bell System Technical Journal, 62*, 1035-1072.

Neal, R. N. and Hinton, G. E. (1993), *A new view of the EM algorithm that justifies incremental and other variants*, University of Toronto, Department of Computer Science preprint.

Nowlan, S.J. (1991). *Soft competitive adaptation: Neural network learning algorithms based on fitting statistical mixtures.* Tech. Rep. CMU-CS-91-126, CMU, Pittsburgh, PA.

Redner, R.A., and Walker, H.F. (1984), Mixture densities, maximum likelihood, and the EM algorithm, *SIAM Review 26*, 195-239.

Titterington, D.M. (1984), Recursive parameter estimation using incomplete data, *J. of Royal Statistical Society, B46*, 257-267.

Tresp, V, Ahmad, S. and Neuneier, R. (1994), Training neural networks with deficient data, *Advances in Neural Information Processing Systems 6*, eds., Cowan, J.D., Tesauro, G., and Alspector, J., San Mateo, CA: Morgan Kaufmann, 128-135.

Waterhouse, S. R., and Robinson, A. J., (1994), Classification using hierarchical mixtures of experts, in *IEEE Workshop on Neural Networks for Signal Processing.*

Wu. C.F. J. (1983), On the convergence properties of the EM algorithm, *The Annals of Statistics, 11*, 95-103.

Xu, L., and Jordan, M.I. (1993a), Unsupervised learning by EM algorithm based on finite mixture of Gaussians, *Proc. of WCNN'93*, Portland, OR, Vol. II, 431-434.

Xu, L., and Jordan, M.I. (1993b), EM learning on a generalized finite mixture model for combining multiple classifiers, *Proc. of WCNN'93*, Portland, OR, Vol. IV, 227-230.

Xu, L., and Jordan, M.I. (1993c), *Theoretical and experimental studies of the EM algorithm for unsupervised learning based on finite Gaussian mixtures*, MIT Computational Cognitive Science, Technical Report 9302, Dept. of Brain and Cognitive Science, MIT, Cambridge, MA.

Xu, L., Jordan, M.I. and Hinton, G.E. (1994), A Modified gating network for the mixtures of experts architecture, *Proc. of WCNN'94*, San Diego, Vol. 2, 405-410.

Yuille, A. L., Stolorz, P. & Utans, J. (1994), Statistical physics, mixtures of distributions and the EM algorithm, *Neural Computation 6*, 334-340.