An Analysis of the EM Algorithm and Entropy-Like Proximal Point Methods[1]

August 5, 2001

Paul Tseng[2]

Abstract

The EM algorithm is a popular method for maximum likelihood estimation from incomplete data. This method may be viewed as a proximal point method for maximizing the log-likelihood function using an integral form of the Kullback-Leibler distance function. Motivated by this interpretation, we consider a proximal point method using an integral form of entropy-like distance function. We give a convergence analysis of the resulting proximal point method in the case where the cluster points lie in the interior of the objective function domain. This result is applied to a normal/independent example and a Gaussian mixture example to establish convergence of the EM algorithm on these examples. Further convergence analysis of the method for maximization over an orthant is given in low dimensions. Sublinear convergence and schemes for accelerating convergence are also discussed.

Key words. EM algorithm, proximal point method, Kullback-Leibler distance, ϕ-divergence, convergence analysis.

1 Introduction

The EM (Expectation-Maximization) algorithm of Dempster, Laird and Rubin [9] is a popular method for maximum likelihood estimation from incomplete data. A recent book on this method [17] writes: "Since its inception in 1977, the EM algorithm has been the subject of intense scrutiny, dozens of applications, numerous extensions, and thousands of publications." Despite its popularity, a rigorous and practical convergence analysis of the EM algorithm is still lacking. The analysis in [9] contains a flaw. A subsequent analysis by Wu [31] gives sufficient conditions for (subsequence) convergence, but it is unclear how to verify the conditions on a given application. Only for the case of computerized tomography with Poisson data has convergence been fully analyzed in [8, 30]; also see [11, page 690] for a survey of subsequent simplifications of this analysis. Convergence rate of the EM algorithm is another issue, as is discussed in [10, 17] and references therein. Horng [10] reportedly gave examples of sublinear convergence, but the analysis has a gap in that it relies on an argument of Ortega and Rheinboldt which is applicable only for linear convergence analysis. In general, convergence and convergence rate analyses are keys to the understanding, effective deployment, and improvement of an iterative method [4, 20].

As we shall see in Section 2, the EM algorithm may be viewed as a proximal point method in the sense of Martinet [15, 16] and Rockafellar [22, 23], but with quadratic

[1]This research is supported by the National Science Foundation Grant CCR-9731273.

[2]Department of Mathematics, University of Washington, Seattle, Washington 98195, U.S.A. (tseng@math.washington.edu)

distance replaced by an integral form of the Kullback-Leibler distance. The Kullback-Leibler distance function can be replaced more generally by entropy-like distances called "ϕ-divergence" [7]–see [12, 26, 27] and references therein for discussions. Accordingly, we generalize in Section 3 the EM algorithm to an inexact proximal point method using an integral form of ϕ-divergence distance. In Section 4, we give sufficient conditions for the iterates generated by this method to be defined, bounded, and for every cluster point to be a stationary point for the objective function (see Thm. 1). An essential feature of this convergence result, distinguishing it from the convergence results in [12, 22, 26, 27], is that it applies to nonconcave maximization problems. This allows the result to be specialized to the EM algorithm with nonconcave log-likelihood function. We also give some lemmas that are very useful for verifying the sufficient conditions. In Section 5, we verify the sufficient conditions on a normal/independent example and a Gaussian mixture example. To our knowledge, this is the first convergence result for the EM algorithm on these well-known examples, thus illustrating the practical utility of the sufficient conditions. Although convergence of proximal point methods for nonconcave maximization (or, equivalently, non-convex minimization) have been studied previously [6, 14, 28], these convergence results assume that the proximity term has certain "distance-like" properties which, as we shall see, either are non-trivial to verify or do not hold for the Kullback-Leibler distance used by the EM algorithm. For example, the continuity and identifiability properties from [6, Assumption 1] are non-trivial to verify for the normal/independent example and the identifiability property does not hold for the Gaussian mixture example. In Section 6, we consider the special case of nonconcave maximization over an orthant, and we analyze convergence of the inexact proximal point method in dimensions 1 and 2. The novelty in the analysis lies in that the solution(s) may lie on the boundary of the orthant, at which the Kullback-Leibler distance is not differentiable. In Section 7, we discuss the issue of sublinear convergence and sketch schemes for accelerating convergence.

In our notation, $\Re^m$ denotes the space of m-dimensional column vectors, and $\Re^m_+$, $\Re^m_{++}$ denote, respectively, the nonnegative and positive orthant in $\Re^m$. For any $S \subset \Re^m$, we denote by int$S$ and ri$S$, respectively, the interior and the relative interior of S. For $\phi \in \Re^m$, ϕ^T denotes the transpose of ϕ and $\|\phi\| = \sqrt{\phi^T \phi}$. We denote $\nabla_1 D(\phi, \theta) = \frac{\partial}{\partial \phi} D(\phi, \theta)$ and use := to mean "define".

2 The EM Algorithm

In this section we describe the maximum likelihood estimation problem and the EM algorithm of [9] for its solution. We postulate that p samples of random variables (x, y) are independently generated from a sample space $X \times Y$ with sampling density $f(x, y|\phi)$ depending on ϕ from a parameter space Φ. The variable y is observed but not x. The observed data $y_1, ..., y_p$ are thus independently generated with marginal density

$$g(y|\phi) := \int_X f(x, y|\phi) dx. \tag{1}$$

To make the integral meaningful, we may assume that X is a Borel subset of $\Re^q$ ($q \geq 1$) with positive Lebesgue measure. The problem is to find $\phi \in \Phi$ that maximizes the

likelihood function

$$\ell(\phi) := \prod_{i=1}^{p} g(y_i|\phi). \tag{2}$$

Notice that f maps $X \times Y \times \Phi$ to $\Re_+$. Also, for convenience we have adopted a slightly different notation from that used in [9].

For a given $\phi^0 \in \Phi$, the EM algorithm generates ϕ^{k+1} from ϕ^k, $k = 0, 1, ...$, according to the iteration (cf. [9, page 6]):

$$\phi^{k+1} \in \arg\max_{\phi\in\Phi} \left\{ Q(\phi, \phi^k) \ := \ \sum_{i=1}^{p} \int_X \ln(f(x, y_i|\phi)) h_i(x|\phi^k) dx \right\}, \tag{3}$$

where $h_i(x|\phi) := f(x, y_i|\phi)/g(y_i|\phi)$. [Here we adopt the convention $Q(\phi, \theta) = -\infty$ whenever the integral is undefined.] Notice that $\int_X h_i(x|\phi)dx = 1$.

It is known [9] and not difficult to see that

$$Q(\phi, \theta) = \ln \ell(\phi) + H(\phi, \theta),$$

where $H(\phi, \theta) := \sum_{i=1}^{p} \int_X \ln(h_i(x|\phi)) h_i(x|\theta)dx$. It is also known that $H(\phi, \theta) \le H(\theta, \theta)$ for all $\phi, \theta \in \Phi$. This suggests the following "distance" function:

$$\begin{aligned} D(\phi, \theta) &:= H(\theta, \theta) - H(\phi, \theta) \\ &= \sum_{i=1}^{p} \int_X -\ln\left(\frac{h_i(x|\phi)}{h_i(x|\theta)}\right) h_i(x|\theta)dx \\ &= \sum_{i=1}^{p} \int_X \psi\left(\frac{h_i(x|\phi)}{h_i(x|\theta)}\right) h_i(x|\theta)dx, \end{aligned} \tag{4}$$

where $\psi(t) := -\ln t + t - 1$ and the last equality also uses the fact $\int_X h_i(x|\phi)dx = 1$. This observation was also made by Chrétien and Hero (see discussions in [6, Section 1]), but with the alternative choice of $\psi(t) = -\ln t$. Both choices of ψ are usable here, though in general we prefer the former since it has bounded lower level sets and other distance-like features [7, 8, 26]. Notice that $D(\phi, \theta)$ is nonnegative for all $\phi, \theta \in \Phi$ and equals zero if $\phi = \theta$. We will see in Lemma 2 that the converse also holds under some mild assumptions on h_i. Using the above observation, we can then rewrite (3) equivalently as

$$\phi^{k+1} \in \arg\max_{\phi\in\Phi} \ \ln(\ell(\phi)) - D(\phi, \phi^k). \tag{5}$$

This is a proximal point iteration in the spirit of Martinet [15, 16] and Rockafellar [22, 23] and extensively studied by others. The novel feature here is that an integral form of the Kullback-Leibler distance function is used.

For the EM algorithm, it was shown in [9] that the generated sequence $\{J(\phi^k)\}$ is nondecreasing and that $\{\phi^k\}$ has certain convergence properties. However, the analysis made assumptions such as $\{\phi^k\}$ converges and $Q(\cdot, \phi^k)$ is strongly concave, which seem difficult to verify. Its proof of convergence of $\{\phi^k\}$ was also incorrect. Subsequently, Wu [31] provided a a more accurate analysis of the EM algorithm using Zangwill's criterion (also see [17, Chap. 3]). However, the proofs in [31] also had gaps and made assumptions that were not stated and/or checked. For example, it assumed throughout that $\{\phi \in \Phi :$

$J(\phi) \geq J(\phi^0)\}$ is a compact subset of intΦ and assumed in [31, Thm. 6] that $\{\|\phi^{k+1} - \phi^k\|\} \to 0$ but did not indicate how to check these assumptions. Also, the proof of [31, Thm. 2] requires $\nabla_1 H$ to be continuous, which was not explicitly stated as an assumption. The sublinear convergence result of Horng [10, Thm. 1.2.1] has a gap in that it asserts a linear convergence result of Ortega and Rheinboldt [20, page 301] extends to sublinear convergence "without any change". This assertion is incorrect since the argument of Ortega and Rheinboldt depends crucially on the algorithmic mapping being a contraction and thus is applicable only for linear convergence analysis.

3 An Entropy-Like Proximal Point Method

Consider the general optimization problem

$$\max_{\phi \in \Phi} J(\phi), \tag{6}$$

where Φ is a nonempty subset of $\Re^m$ ($m \geq 1$) and $J : \Phi \to \Re$. The proximal point interpretation of the EM algorithm (5), as well as previous works on proximal point methods, suggest the following general proximal point method whereby, for a given $\phi^0 \in \Phi$, we generate ϕ^{k+1} from ϕ^k, $k = 0, 1, ...$, according to:

$$\phi^{k+1} \in \arg\max_{\phi \in \Phi} J(\phi) - D(\phi, \phi^k), \tag{7}$$

where $D : \Phi \times \Phi \to \Re_+$ satisfies $D(\phi, \phi) = 0$ for all $\phi \in \Phi$ and the maximum is assumed to be attained. Proximal point methods have been much studied, beginning with the seminal works of Martinet [15, 16] and Rockafellar [22, 23] using a quadratic proximity term. Recent works on methods using non-quadratic proximity terms are described in [2, 3, 5, 6, 12, 13, 14, 19, 25, 26, 27, 29] and references therein. The above proximal point iteration (7) can be further generalized by scaling D with a stepsize/relaxation parameter $\lambda_k > 0$. Our convergence results readily extend to this more general method provided that $\liminf_{k\to\infty} \lambda_k > 0$ and $\limsup_{k\to\infty} \lambda_k < \infty$. However, in the case of the EM algorithm where D has the form (4), using $\lambda_k \neq 1$ destroys the simple form of the iteration, leading to signficant increases in the work per iteration. For this reason and for simplicity, we will not further consider this more general method.

Since J is not concave, computing an (inexact) local/global maximum for (7) may be difficult and expensive in general. On the other hand, if $J(\cdot) - D(\cdot, \phi^k)$ is differentiable, as often is the case with applications of the EM algorithm (see Section 5 for examples), then any local/global maximum of (7) is a stationary point of this function and its function value is no less than the function value of ϕ^k. Computing such a stationary point inexactly is much easier and, as we will see, suffices for establishing convergence to a stationary point of (6). Specifically, we consider the following inexact proximal point method whereby $\phi^{k+1} \in \Phi$ is chosen to satisfy

$$\|\nabla J(\phi^{k+1}) - \nabla_1 D(\phi^{k+1}, \phi^k)\| \leq \delta_k, \quad J(\phi^{k+1}) - D(\phi^{k+1}, \phi^k) \geq J(\phi^k), \tag{8}$$

with $\delta_k \in \Re_+$ a tolerance parameter tending to zero as $k \to \infty$. In many applications of the EM algorithm, the maximum can be computed exactly, so that $\delta_k = 0$. In general,

for any $\delta_k > 0$, a ϕ^{k+1} satisfying (8) can be computed in a finite number of steps using a feasible descent method starting at ϕ^k. Such methods are described in, e.g., [4] and references therein. We note that

The special form (4) of D used for the EM algorithm motivates an integral form of D based on ϕ-divergence, which we describe below. Special properties of this entropy-like distance will play an essential role when we apply the general convergence theory for (8) to derive convergence of the EM algorithm on specific examples. In what follows, we denote by Ψ the class of strictly convex, continuously differentiable functions $\psi : \Re_{++} \to \Re$ satisfying $\lim_{t\to 0^+} \psi'(t) = -\infty$, $\psi(1) = \psi'(1) = 0$ for all $t > 0$ (cf. [12, 27]). Thus each $\psi \in \Psi$ satisfies $\psi(t) \geq 0$ for all $t > 0$ and $\psi(t) = 0$ if and only if $t = 1$. As is discussed in [12, 27], examples of ψ include:

$$\begin{array}{ll} \psi(t) & = -\ln t + t - 1. \\ \psi(t) & = t \ln t - t + 1. \\ \psi(t) & = (\sqrt{t} - 1)^2. \\ \psi(t) & = t + t^{-1} - 2. \end{array} \tag{9}$$

Let X be a Borel subset of $\Re^q$ with positive Lebesgue measure. Let $\psi \in \Psi$ and $h_i : X \times \Phi \to \Re_{++}$ be such that $\int_X h_i(x|\phi)dx = 1$ for all $\phi \in \Phi$, $i = 1, ..., p$, and

$$D(\phi, \theta) \ := \ \sum_{i=1}^{p} \int_X \psi\left(\frac{h_i(x|\phi)}{h_i(x|\theta)}\right) h_i(x|\theta)dx \tag{10}$$

is defined (i.e., integrand is Lebesgue integrable on X [1]) and finite for all $\phi, \theta \in \mathrm{ri}\Phi$. [We set $D(\phi, \theta) = -\infty$ whenever an integral in (10) is undefined.] Let $\mathcal{H}[X, \psi]$ denote the set of such h_i.

By our assumptions on h and ψ, D is nonnegative-valued and $D(\phi, \phi) = 0$ for all $\phi \in \mathrm{ri}\Phi$. With this property of D, we may alternatively view (7) as a block coordinate ascent method whereby $J(\phi) - D(\phi, \theta)$ is alternately maximized with respect to ϕ (with θ fixed) and with respect to θ (with ϕ fixed). This interpretation differs from that expressed in [8, 18], which view the EM algorithm as alternately maximizing a joint function of the parameters (M step) and of the distribution over the unobserved variables (E step). The proximity term (10) in the parameter space is a bit unusual and, to our knowledge, has not been much studied beyond [8]. We will study its differential/continuity properties in the case where ψ and h_i have the special form (11) and (12) below.

It is readily seen that (5) is a special case of (7), (10) with

$$J(\phi) = \ln(\ell(\phi)), \qquad \psi(t) = -\ln(t) + t - 1, \qquad h_i(x|\phi) = f(x, y_i|\phi)/g(y_i|\phi). \tag{11}$$

It was also observed in [9] that if f has the exponential form

$$f(x, y_i|\phi) = a_i(\phi)b_i(x)e^{-c_i(\phi)^T d_i(x)}, \tag{12}$$

for some $a_i : \Phi \to \Re_{++}$, $b_i : X \to \Re_{++}$, $c_i : \Phi \to \Re^q_{++}$, $d_i : X \to \Re^q_{++}$ (also see [17, page 26]), then

$$Q(\phi, \theta) = \sum_{i=1}^{p} \ln(a_i(\phi)) + \int_X \ln(b_i(x))h_i(x|\theta)dx - c_i(\phi)^T \left(\int_X d_i(x)f(x, y_i|\theta)dx\right)/g(y_i|\theta).$$

The second term from the right does not depend on ϕ and hence can be ignored. The last term on the right requires 1 integration to evaluate $g(y_i|\theta)$ and q integrations to evaluate the integral inside the parentheses. Summing over all $i = 1, ..., p$, we see that $p(1+q)$ integrations are needed to compute the coefficients multiplying the components of $c_i(\phi)$, $i = 1, ..., p$. If $c_i(\cdot)$ and $\ln(a_i(\cdot))$ have simple forms, then the maximization with respect to ϕ would be relatively cheap. Thus, the work per iteration (3) is not sensitive to m, the number of parameters. In contrast, if we use a general optimization method to maximize J, such as derivative-free method or finite-difference gradient method [4], then more than m function evaluations of J may be required before an ascent in J is obtained. Each function evaluaton requires p integrations, so the total work is roughly mp. Thus, when $q < m$, the EM algorithm may require fewer integral evaluations per ascent in J. However, this feature of the EM algorithm is very special and does not appear to carry beyond the choice (11) of J, ψ, and h_i.

4 Convergence of the Proximal Point Method

In this section we analyze the convergence properties of the inexact proximal point method (8), (10) in the case where the optimal solution(s) of (6) lies in the interior of Φ. In particular, we give conditions on J and D under which the sequence $\{\phi^k\}$ is defined, bounded and every cluster point of which is a stationary point of J (see Thm. 1 and Lemmas 1, 2, 3).

We begin with a general convergence result for the inexact proximal point method (8).

Theorem 1 *Let* $\Phi^0 := \{\phi \in \Phi : J(\phi) \geq J(\phi^0)\}$. *Assume*

A1: ∇J, D *and* $\nabla_1 D$ *are defined and continuous on, respectively,* Φ^0, $\Phi^0 \times \Phi^0$ *and* $\Phi^0 \times \Phi^0$,

where $D : \Phi \times \Phi \to \Re_+$ *satisfies* $D(\phi, \phi) = 0$ *for all* $\phi \in \Phi$. *Also, assume* $\phi^{k+1} \in \Phi$ *satisfying* (7), *with* $\delta_k \to 0$, *is defined for* $k = 0, 1, ...$ *Then* $\{\phi^k\}$ *has the following properties:*

(a) $J(\phi^{k+1}) \geq J(\phi^k)$ *for all* k.

(b) $\phi^k \in \Phi^0$ *for all* k.

(c) *If* Φ^0 *is closed and* $\{\phi^{k+1} - \phi^k\} \to 0$, *then every cluster point* ϕ^∞ *of* $\{\phi^k\}$ *satisfies* $\nabla J(\phi^\infty) = 0$.

Proof. By (8) and $D(\phi^{k+1}, \phi^k) \geq 0$, we obtain $J(\phi^{k+1}) \geq J(\phi^k)$ for $k = 0, 1, ...$ and hence $J(\phi^k) \geq J(\phi^0)$ for $k = 0, 1, ...$ Since $\phi^k \in \Phi$, then $\phi^k \in \Phi^0$ for $k = 0, 1, ...$

Assume Φ^0 is closed and $\{\phi^{k+1} - \phi^k\} \to 0$. Let ϕ^∞ be any cluster point of $\{\phi^k\}$, if it exists. Then there is some subsequence $\{\phi^k\}_{k \in K}$ ($K \subseteq \{0, 1, ...\}$) converging to ϕ^∞. Since $\{\phi^{k+1} - \phi^k\} \to 0$, then $\{\phi^{k+1}\}_{k \in K} \to \phi^\infty$. Then (8) and $\delta_k \to 0$ imply

$$\nabla J(\phi^{k+1}) - \nabla_1 D(\phi^{k+1}, \phi^k) \to 0 \quad \text{as } k \to \infty.$$

Since $\phi^k, \phi^{k+1} \in \Phi^0$ and Φ^0 is closed, then $\phi^\infty \in \Phi^0$. The continuity of ∇J and $\nabla_1 D$ on, respectively, Φ^0 and $\Phi^0 \times \Phi^0$ yield in the limit

$$\nabla J(\phi^\infty) - \nabla_1 D(\phi^\infty, \phi^\infty) = 0.$$

By assumption, $D(\cdot, \phi^\infty)$ is minimized at ϕ^∞, so $\nabla_1 D(\phi^\infty, \phi^\infty) = 0$. Thus, $\nabla J(\phi^\infty) = 0$. ■

We note that Thm. 1(a), (b) was shown in [9] for the case of the EM algorithm, i.e., with D given by (10) and J, ψ, h given by (11). If J is not continuously differentiable but has outer semicontinuous subgradients ∂J [24, Chaps. 8, 10], then (8) and Thm. 1(c) may be accordingly extended to yield $0 \in \partial J(\phi^\infty)$. Related convergence result are given in [6, 28] and references therein. In applications of the EM algorithm, however, J is typically differentiable. If J is of the form $J(\phi) = \tilde{J}(A\phi)$, where $\tilde{J}$ is locally strongly concave and A is a linear mapping (or, more generally, a local error bound for J holds), then, by using Lemma 2, (8), and [14, Thms. 2.1, 3.1], Thm. 1(c) can be extended to obtain a (local) linear convergence rate result for $\{\phi^k\}$ and $\{J(\phi^k)\}$. However, verifying the above assumption a priori seems difficult.

The following lemma gives sufficient condition on J for $\{\phi^k\}$ generated by (8) to be defined, bounded, and satisfies $\{\phi^{k+1} - \phi^k\} \to 0$. This lemma and two subsequent lemmas will be used in the convergence analysis of the EM algorithm in Section 5.

Lemma 1 *Assume, in addition to assumption A1, that*

A2: Φ^0 *is a compact subset of* $\mathrm{int}\Phi$.

Then ϕ^{k+1} *satisfying* (8) *is defined for any* $\delta_k \geq 0$, $k = 0, 1, ...$, *and* $\{\phi^k\}$ *is bounded. If*

A3: $D(\phi, \theta) > 0$ *for all* $\theta \neq \phi \in \Phi^0$,

then $\{\phi^{k+1} - \phi^k\} \to 0$.

Proof. We prove that ϕ^{k+1} is defined by induction on k. Clearly ϕ^0 is defined. Suppose ϕ^k is defined for some $k \geq 0$. Since $J(\cdot) - D(\cdot, \phi^k)$ is continuous on Φ^0 and Φ^0 is compact, then it attains its global maximum over Φ at some point $\bar{\phi} \in \Phi^0 \subseteq \mathrm{int}\Phi$. Since our assumption implies the gradient of this function is defined and continuous at $\bar{\phi}$, then

$$\nabla J(\bar{\phi}) - \nabla_1 D(\bar{\phi}, \phi^k) = 0, \quad J(\bar{\phi}) - D(\bar{\phi}, \phi^k) \geq J(\phi^k).$$

Then, taking $\phi^{k+1} = \bar{\phi}$ would satisfy (8) for any $\delta_k \geq 0$, so ϕ^{k+1} is defined. By induction, ϕ^{k+1} is defined for $k = 0, 1, ...$

By Thm. 1(b), $\phi^k \in \Phi^0$ for all k. Since Φ^0 is compact, then $\{\phi^k\}$ is bounded.

Assume in addition that A3 holds. Suppose $\{\phi^{k+1} - \phi^k\} \not\to 0$. Then there would exist a subsequence $\{\phi^k\}_{k \in K}$ ($K \subseteq \{0, 1, ...\}$) such that $\{\phi^{k+1} - \phi^k\}_{k \in K} \not\to 0$. Since $\{\phi^k\}$ is bounded, by passing to a subsequence if necessary, we can assume that $\{\phi^k\}_{k \in K}$ converges to some ϕ^∞ and $\{\phi^{k+1}\}_{k \in K}$ converges to some $\bar{\phi} \neq \phi^\infty$. Since Φ^0 is closed, $\phi^\infty, \bar{\phi} \in \Phi^0$. Also, by Thm. 1(a), $\{J(\phi^k)\}$ is nondecreasing. Since J is continuous on Φ^0, then $\lim_{k \to \infty} J(\phi^k) = J(\phi^\infty)$ and, by (8),

$$\{D(\phi^{k+1}, \phi^k)\} \to 0.$$

Since D is continuous on $\Phi^0 \times \Phi^0$, this implies $D(\bar{\phi}, \phi^\infty) = 0$. Since $\bar{\phi}, \phi^\infty \in \Phi^0$, A3 implies $\bar{\phi} = \phi^\infty$, a contradiction. Thus, $\{\phi^{k+1} - \phi^k\} \to 0$. ■

Assumption A2 was also used in [31] for the EM algorithm while A2 and A3 were used in [6] for a general proximal point method. Under A3, we have that $D(\phi, \cdot)$ is minimized at ϕ. Thus (8) may be viewed as a block-coordinate ascent method, whereby $J(\phi) - D(\phi, \theta)$ is alternately maximized (locally) with respect to ϕ (with θ held fixed) and with respect to θ (with ϕ held fixed). The following lemma gives sufficient conditions for assumption A3 to be satisfied when D has the form (10). Under these conditions, each maximization is attained at a unique point.

Lemma 2 *Let $\psi \in \Psi$ and let $h_i \in \mathcal{H}[X, \psi]$, $i = 1, ..., p$. Define D by (10). For any $\phi, \theta \in \Phi$, if $h_{\bar{i}}(x|\phi) \neq h_{\bar{i}}(x|\theta)$ for some $\bar{i} \in \{1, ..., p\}$ and some $x \in \text{int}X$ and both $h_{\bar{i}}(\cdot, \phi)$ and $h_{\bar{i}}(\cdot, \theta)$ are continuous at x, then $D(\phi, \theta) > 0$.*

Proof. Since $h_{\bar{i}}(x|\phi) \neq h_{\bar{i}}(x|\theta)$ and both $h_{\bar{i}}(\cdot, \phi)$ and $h_{\bar{i}}(\cdot, \theta)$ are continuous at x, there exist positive scalars c_1, c_2 and an open subset $O \subset X$ such that

$$|h_{\bar{i}}(x|\phi)/h_{\bar{i}}(x|\theta) - 1| \geq c_1, \quad h_{\bar{i}}(x|\theta) \geq c_2 \qquad \forall x \in O.$$

Since ψ is convex continuous and nonnegative-valued on $\Re_{++}$ and $\psi(t) = 0$ if and only if $t = 1$, there exists a positive scalar such that $\psi(t) \geq c_3$ whenever $|t - 1| \geq c_1$. These relations, together with h_i being positive-valued, imply

$$D(\phi, \theta) \;\geq\; \int_O \psi\left(\frac{h_{\bar{i}}(x|\phi)}{h_{\bar{i}}(x|\theta)}\right) h_{\bar{i}}(x|\theta) dx \;\geq\; \int_O c_3 c_2 dx \;=\; c_3 c_2 \left(\int_O dx\right) \;>\; 0,$$

where the last inequality follows from O having positive Lebesgue measure. ■

Below we verify assumption A1 for the case where J, ψ, h are of the form (11) with ℓ, g, f of the form (1), (2), (12).

Lemma 3 *Assume f is of the form (12), with $a_i : \Phi \to \Re_{++}$, $c_i : \Phi \to \Re^q_{++}$ and Borel measurable $b_i : X \to \Re_{++}$, $d_i : X \to \Re^q_{++}$, $i = 1, ..., p$. Assume $\underline{\Phi}$ is a nonempty open convex subset of Φ over which a_i and $c_i = (c_{ij})_{j=1}^q$ are continuously differentiable and such that*

$$\int_X b_i(x) e^{-\underline{c}_i^T d_i(x)} \left((1 + |\ln(b_i(x))|) \left(1 + \sum_{j=1}^q d_{ij}(x)\right) + \left(\sum_{j=1}^q d_{ij}(x)\right)^2 \right) dx \;<\; \infty \forall i, \quad (13)$$

where $d_i = (d_{ij})_{j=1}^q$, $\underline{c}_i = (\underline{c}_{ij})_{j=1}^q$, with $\underline{c}_{ij} := \inf_{\phi \in \underline{\Phi}} c_{ij}(\phi)$. Let D and J be given by (10), (11) with ℓ, g of the form (1), (2). Then ∇J is defined and continuous on $\underline{\Phi}$, while D and $\nabla_1 D$ are defined and continuous on $\underline{\Phi} \times \underline{\Phi}$.

Proof. Here J and D are sums of p terms having the same form, so it suffices to consider the case of $p = 1$. In this case, there is only one sample of y, so we can drop the subscript i and abbreviate $f(x, y|\phi)$ and $g(y|\phi)$ as $f(x|\phi)$ and $g(\phi)$ with no loss of ambiguity.

We have from (1), (2), (12) and X having positive Lebesgue measure that

$$\ell(\phi) = g(\phi) = a(\phi) \int_X b(x) e^{-c(\phi)^T d(x)} dx > 0.$$

Since a is continuously differentiable and $J(\phi) = \ln(\ell(\phi))$, to show that g and J are continuously differentiable on $\underline{\Phi}$, it suffices to show that

$$w(\phi) := \int_X b(x) e^{-c(\phi)^T d(x)} dx$$

is continuously differentiable on $\underline{\Phi}$. To show this, fix any $\phi \in \underline{\Phi}$. For any $\epsilon > 0$, there exists $\delta > 0$ such that $\tilde{\phi} \in \underline{\Phi}$ and $\|\tilde{\phi} - \phi\| < \delta$ imply $|c_j(\tilde{\phi}) - c_j(\phi)| < \epsilon$ and $\|\nabla c_j(\tilde{\phi}) - \nabla c_j(\phi)\| < \epsilon$ for all i. Then, for any such $\tilde{\phi}$, the mean value theorem yields

$$\begin{aligned}
r &:= w(\tilde{\phi}) - w(\phi) - \left(- \int_X b(x) e^{-c(\phi)^T d(x)} \sum_{j=1}^{q} \nabla c_j(\phi)^T (\tilde{\phi} - \phi) d_j(x) dx \right) \\
&= \int_X b(x) \left(e^{-c(\tilde{\phi})^T d(x)} - e^{-c(\phi)^T d(x)} + e^{-c(\phi)^T d(x)} \sum_{j=1}^{q} \nabla c_j(\phi)^T (\tilde{\phi} - \phi) d_j(x) \right) dx \\
&= \int_X b(x) \left(-e^{-c(\varphi(x))^T d(x)} \sum_{j=1}^{q} \nabla c_j(\varphi(x))^T (\tilde{\phi} - \phi) d_j(x) \right. \\
&+ \left. e^{-c(\phi)^T d(x)} \sum_{j=1}^{q} \nabla c_j(\phi)^T (\tilde{\phi} - \phi) d_j(x) \right) dx \\
&= \int_X b(x) \left(-e^{-c(\varphi(x))^T d(x)} \sum_{j=1}^{q} (\nabla c_j(\varphi(x)) - \nabla c_j(\phi))^T (\tilde{\phi} - \phi) d_j(x) \right. \\
&+ \left. \left(e^{-c(\phi)^T d(x)} - e^{-c(\varphi(x))^T d(x)} \right) \sum_{j=1}^{q} \nabla c_j(\phi)^T (\tilde{\phi} - \phi) d_j(x) \right) dx \\
&= \int_X b(x) \left(-e^{-c(\varphi(x))^T d(x)} \sum_{j=1}^{q} (\nabla c_j(\varphi(x)) - \nabla c_j(\phi))^T (\tilde{\phi} - \phi) d_j(x) \right. \\
&+ \left. e^{-\gamma(x)} \left(-c(\phi) + c(\varphi(x)) \right)^T d(x) \sum_{j=1}^{q} \nabla c_j(\phi)^T (\tilde{\phi} - \phi) d_j(x) \right) dx,
\end{aligned}$$

where $\varphi(x)$ is on the line segment between ϕ and $\tilde{\phi}$ and $\gamma(x)$ is between $c(\phi)^T d(x)$ and $c(\varphi(x))^T d(x)$. Since $\underline{\Phi}$ is convex, both ϕ and $\varphi(x)$ are in $\underline{\Phi}$ and so $c_j(\phi) \geq \underline{c}_j$ and $c_j(\varphi(x)) \geq \underline{c}_j$. This and $d(x) > 0$ imply $c(\varphi(x))^T d(x) \geq \underline{c}^T d(x)$ and $\gamma(x) \geq \underline{c}^T d(x)$. We also have $\|\varphi(x) - \phi\| < \delta$, so that $|c_j(\phi) - c_j(\varphi(x))| \leq \epsilon$ and $\|\nabla c_j(\varphi(x)) - \nabla c_j(\phi)\| \leq \epsilon$ for all i. Thus

$$\begin{aligned}
|r| &\leq \int_X b(x) \left(e^{-\underline{c}^T d(x)} \left(\sum_{j=1}^{q} d_j(x) \right) \max_j |(\nabla c_j(\varphi(x)) - \nabla c_j(\phi))^T (\tilde{\phi} - \phi)| \right. \\
&+ \left. e^{-\underline{c}^T d(x)} \left(\sum_{j=1}^{q} d_j(x) \right)^2 \max_j |c_j(\phi) - c_j(\varphi(x))| \cdot \max_j |\nabla c_j(\phi)^T (\tilde{\phi} - \phi)| \right) dx \\
&\leq \left(\int_X b(x) e^{-\underline{c}^T d(x)} \left(\sum_{j=1}^{q} d_j(x) + \left(\sum_{j=1}^{q} d_j(x) \right)^2 \right) dx \right) \epsilon (1 + \max_j \|\nabla c_j(\phi)\|) \|\tilde{\phi} - \phi\|.
\end{aligned}$$

This shows that w is Fréchet-differentiable at each $\phi \in \underline{\Phi}$ with

$$\nabla w(\phi) = -\int_X b(x) e^{-c(\phi)^T d(x)} \sum_{j=1}^{q} \nabla c_j(\phi) d_j(x) dx.$$

By (13), this integral is finite.

To show that ∇w is continuous on $\underline{\Phi}$, we note that, for any $\phi, \tilde{\phi} \in \underline{\Phi}$, the mean value theorem yields

$$\begin{aligned}
& \nabla w(\tilde{\phi}) - \nabla w(\phi) \\
= & \int_X b(x) \left(e^{-c(\phi)^T d(x)} \sum_{j=1}^{q} \nabla c_j(\phi) - e^{-c(\tilde{\phi})^T d(x)} \sum_{j=1}^{q} \nabla c_j(\tilde{\phi}) \right) d_j(x) dx \\
= & \int_X b(x) \left(\left(e^{-c(\phi)^T d(x)} - e^{-c(\tilde{\phi})^T d(x)} \right) \sum_{j=1}^{q} \nabla c_j(\phi) + e^{-c(\tilde{\phi})^T d(x)} \sum_{j=1}^{q} (\nabla c_j(\phi) - \nabla c_j(\tilde{\phi})) \right) d_j(x) dx \\
= & \int_X b(x) \left(e^{-\gamma(x)} (-c(\phi) + c(\tilde{\phi}))^T d(x) \sum_{j=1}^{q} \nabla c_j(\phi) + e^{-c(\tilde{\phi})^T d(x)} \sum_{j=1}^{q} (\nabla c_j(\phi) - \nabla c_j(\tilde{\phi})) \right) d_j(x) dx,
\end{aligned}$$

where $\gamma(x)$ is between $c(\phi)^T d(x)$ and $c(\tilde{\phi})^T d(x)$. Since $\phi, \tilde{\phi} \in \underline{\Phi}$, we have $c_j(\phi) \geq \underline{c}_j$ and $c_j(\tilde{\phi}) \geq \underline{c}_j$. This and $d(x) > 0$ imply $\gamma(x) \geq \underline{c}^T d(x)$ and $c(\tilde{\phi})^T d(x) \geq \underline{c}^T d(x)$. Thus,

$$\begin{aligned}
\|\nabla w(\tilde{\phi}) - \nabla w(\phi)\| \leq & \int_X b(x) \left(e^{-\underline{c}^T d(x)} \max_j |c_j(\phi) - c_j(\tilde{\phi})| \left(\sum_{j=1}^{q} d_j(x) \right)^2 \max_j \|\nabla c_j(\phi)\| \right. \\
& + \left. e^{-\underline{c}^T d(x)} \sum_{j=1}^{q} d_j(x) \max_j \|\nabla c_j(\phi) - \nabla c_j(\tilde{\phi})\| \right) dx \\
\leq & \left(\int_X b(x) e^{-\underline{c}^T d(x)} \left(\left(\sum_{j=1}^{q} d_j(x) \right)^2 + \sum_{j=1}^{q} d_j(x) \right) dx \right) \\
& \cdot \left(\max_j |c_j(\phi) - c_j(\tilde{\phi})| \cdot \max_j \|\nabla c_j(\phi)\| + \max_j \|\nabla c_j(\phi) - \nabla c_j(\tilde{\phi})\| \right).
\end{aligned}$$

Since, by (13), the integral on the right-hand side is finite and c_j and ∇c_j are continuous on $\underline{\Phi}$, this shows that ∇w is continuous on $\underline{\Phi}$.

Finally, letting H be defined as in Section 2, we have from (1), (2), (11) and (12) that

$$\begin{aligned}
H(\phi, \theta) = & \frac{1}{g(\theta)} \left(\int_X \ln(f(x|\phi)) f(x|\theta) dx \right) - \ln(g(\phi)) \\
= & \frac{a(\theta)}{g(\theta)} \left(\int_X \left(\ln(a(\phi)) + \ln(b(x)) - c(\phi)^T d(x) \right) b(x) e^{-c(\theta)^T d(x)} dx \right) - \ln(g(\phi)) \\
= & \frac{a(\theta)}{g(\theta)} \left(\ln(a(\phi)) \int_X b(x) e^{-c(\theta)^T d(x)} dx + \int_X \ln(b(x)) b(x) e^{-c(\theta)^T d(x)} dx \right. \\
& - \left. c(\phi)^T \int_X d(x) b(x) e^{-c(\theta)^T d(x)} dx \right) - \ln(g(\phi)).
\end{aligned}$$

Since $D(\phi,\theta) = H(\theta,\theta) - H(\phi,\theta)$, to show that D and $\nabla_1 D$ are defined and continuous on $\underline{\Phi} \times \underline{\Phi}$, it suffices to show that the three integrals on the right-hand side are defined and continuous in $\theta \in \underline{\Phi}$. This can be argued similarly as in the argument showing that ∇w is continuous on $\underline{\Phi}$. ■

Although we have considered only the finite-dimensional setting of $\Phi = \Re^m$, the inexact proximal point method (8), (10) can also be applied to the infinite-dimensional setting of, say, Φ being a real Hilbert space. However, the analysis becomes more complicated since we then need to distinguish between weak and strong topology when dealing with convergence and continuity issues.

5 Case Studies of Convergence for the EM Algorithm

In this section we illustrate how Thm. 1 and Lemmas 1, 2, 3 can be applied to analyze convergence of the EM algorithm on a normal/independent example and a Gaussian mixture example [9, 17]. We will derive simple conditions on the data and the initial parameter values under which parameter values generated by the EM algorithm are asymptotically stationary points of the likelihood function. To our knowledge, such convergence results are new for these two well-known applications of the EM algorithm.

Example 1 [9, page 19]. Here, the observed data $y_1, ..., y_p \in \Re$ are assumed to be an i.i.d. random sample from a population such that $(y_i - \mu)/\sqrt{\vartheta/x_i}$ has a $N(0,1)$ distribution conditioned on x_i, and unobserved variables $x_1, ..., x_p$ form an i.i.d. random sample from the density $(\beta^\alpha/\Gamma(\alpha))x_i^{\alpha-1}e^{-\beta x_i}$ $(\alpha > 0, \beta > 0)$. The parameter set $\phi = (\mu, \vartheta)$ is to be estimated. Then $X = \Re_{++}$, $\Phi = \Re \times \Re_{++}$, and

$$f(x, y|\phi) = \frac{e^{-(y-\mu)^2 x/(2\vartheta)}}{\sqrt{2\pi\vartheta/x}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} = \frac{\beta^\alpha}{\sqrt{2\vartheta}\Gamma(\alpha)} x^{\alpha-1/2} e^{-\left(\beta + (y-\mu)^2/(2\vartheta)\right)x}.$$

Thus, $f(x, y_i|\phi)$ has the form (12) with $a_i(\phi) = 1/\sqrt{2\vartheta}$, $b_i(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1/2}$, $c_i(\phi) = \beta + (y_i - \mu)^2/(2\vartheta)$, $d_i(x) = x$. In what follows, g and ℓ are given by (1), (2), J, ψ, h_i are given by (11), and D is given by (10).

It is readily verified that the assumption of Lemma 3 holds, with $\underline{\Phi} = \Phi$ and $\underline{c_i} = \beta$. Hence, by Lemma 3, ∇J is continuous on Φ while D and $\nabla_1 D$ are continuous on $\Phi \times \Phi$. Also, (1) and (2) yield

$$\begin{aligned}
\ell(\phi) &= \prod_{i=1}^p \int_0^\infty \frac{\beta^\alpha}{\sqrt{2\pi\vartheta}\Gamma(\alpha)} x^{\alpha-1/2} e^{-c_i(\phi)x} dx \\
&= \prod_{i=1}^p \frac{\beta^\alpha}{\sqrt{2\pi\vartheta}\Gamma(\alpha)} \frac{\Gamma(\alpha+1/2)}{(c_i(\phi))^{\alpha+1/2}} \\
&= \prod_{i=1}^p \frac{\beta^\alpha \Gamma(\alpha+1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \left(\beta\vartheta + (y_i-\mu)^2/2\right)^{-1/2} (\beta + (y_i-\mu)^2/(2\vartheta))^{-\alpha}. \qquad (14)
\end{aligned}$$

It can be seen using (14) that J is not a concave function. As $\vartheta + |\mu| \to \infty$, we have $\beta\vartheta + (y_i - \mu)^2/2 \to \infty$ and hence (14) yields $\ell(\phi) \to 0$. Thus

$$\Phi^0 := \{\phi \in \Phi : \ell(\phi) \geq \ell(\phi^0)\} = \{\phi \in \Phi : J(\phi) \geq J(\phi^0)\}$$

is bounded for any $\phi^0 \in \Phi$ (since $\ell(\phi^0) > 0$).

Next, we study when Φ^0 is closed. Let $\bar{p}$ be the largest number of identical elements of $\{y_1, ..., y_p\}$. Then there exists $I \subset \{1, ..., p\}$ of cardinality $\bar{p}$ such that $y_i = y_j$ for all $i, j \in I$. If $\bar{p}/2 > \alpha(p - \bar{p})$, then setting $\mu = y_i$ for $i \in I$ and letting $\vartheta \to 0$, we see from (14) that $\ell(\phi) \to \infty$ at the asymptotic rate of $\vartheta^{\alpha(p-\bar{p})-\bar{p}/2}$. Thus Φ^0 is not closed for any $\phi^0 \in \Phi$. If $\bar{p}/2 = \alpha(p - \bar{p})$, then $\ell(\phi)$ instead tends to a positive limit independent of ϕ^0. So, for $\phi^0 \in \Phi$ with $\ell(\phi^0)$ sufficiently near 0 (such ϕ^0 can be obtained by taking $\vartheta + |\mu|$ sufficiently large), Φ^0 is not closed. If $\bar{p}/2 < \alpha(p - \bar{p})$, then there exists $c > 0$ such that, for any $\phi = (\mu, \vartheta)$, there exists $J \subset \{1, ..., p\}$ of cardinality at least $p - \bar{p}$ such that $(y_i - \mu)^2/2 \geq c$ for $i \in J$. For $i \notin J$, we have trivially $(y_i - \mu)^2/2 \geq 0$. Then, (14) yields

$$\begin{aligned} \ell(\phi) &\leq \left(\frac{\beta^\alpha \Gamma(\alpha + 1/2)}{\sqrt{2\pi}\Gamma(\alpha)} \right)^p (\beta\vartheta + c)^{-(p-\bar{p})/2} (\beta\vartheta)^{-\bar{p}/2} (\beta + c/\vartheta)^{-\alpha(p-\bar{p})} \beta^{-\alpha\bar{p}} \\ &= O(\vartheta^{\alpha(p-\bar{p})-\bar{p}/2}) \to 0 \quad \text{as} \quad \vartheta \to 0. \end{aligned}$$

This shows that if $(\mu, \vartheta) \in \Phi^0$, then ϑ cannot tend to zero. Thus Φ^0 is closed for any $\phi^0 \in \Phi$.

Finally, we have from (11) that

$$h_i(x|\phi) = x^{\alpha-1/2} e^{-c_i(\phi)x} \frac{(c_i(\phi))^{\alpha+1/2}}{\Gamma(\alpha + 1/2)} = \frac{x^{\alpha-1/2}}{\Gamma(\alpha + 1/2)} e^{-c_i(\phi)x + (\alpha+1/2)\ln(c_i(\phi))}.$$

Notice that $h_i(x|\phi)$ depends on ϕ through $c_i(\phi)$ only. As a byproduct of D being continuous on $\Phi \times \Phi$, we have that $h_i \in \mathcal{H}[X, \psi]$. Clearly $h_i(\cdot, \phi)$ is continuous at every $x \in \text{int}X = X$. Assume that $\{y_1, ..., y_p\}$ has at least three distinct elements (so $p \geq 3$). We show below that $h_i(x|\phi) \neq h_i(x|\theta)$ for some $x \in X$ whenever $\phi \neq \theta$. Then, by Lemma 2, assumption A3 holds.

Fix any $\phi = (\mu, \vartheta)$ and $\theta = (\tilde{\mu}, \tilde{\vartheta})$ in Φ. Let $\beta_i := c_i(\phi)$, $\tilde{\beta}_i := c_i(\theta)$ for $i = 1, \ldots, p$. Consider any $x_i \in X$ such that $x_i \geq (\alpha + 1/2)/\beta_i + C_i/2$ if $\tilde{\beta}_i \geq \beta_i$ and $x_i \leq \alpha C_i$ else, where $C_i := 1/\max_i\{\beta_i, \tilde{\beta}_i\}$. Then the mean value theorem yields

$$\begin{aligned} & |(-\beta_i x_i + (\alpha + 1/2)\ln(\beta_i)) - (-\tilde{\beta}_i x_i + (\alpha + 1/2)\ln(\tilde{\beta}_i))| \\ = \; & |(\tilde{\beta}_i - \beta_i)x_i + (\alpha + 1/2)(\beta_i - \tilde{\beta}_i)/\hat{\beta}_i| \\ = \; & |(\tilde{\beta}_i - \beta_i)(x_i - (\alpha + 1/2)/\hat{\beta}_i)| \\ \geq \; & |\tilde{\beta}_i - \beta_i| C_i/2, \end{aligned}$$

where $\hat{\beta}_i$ is between β_i and $\tilde{\beta}_i$, and the last inequality follows from our choice of x and the fact $C_i \leq 1/\hat{\beta}_i$. From the form of $h_i(x|\phi)$ shown above, we see that this implies $h_i(x|\phi) \neq h_i(x|\theta)$ whenever $\tilde{\beta}_i \neq \beta_i$. Thus, it remains to show that $\tilde{\beta}_i \neq \beta_i$ for some i whenever $\phi \neq \theta$. Suppose $\tilde{\beta}_i = \beta_i$ for all i. We divide our argument into two cases. First, suppose $|y_i - \mu| = |y_i - \tilde{\mu}|$ for some i. Then, since $\tilde{\beta}_i = \beta_i$, we obtain $\vartheta = \tilde{\vartheta}$. This in turn implies $|y_i - \mu| = |y_i - \tilde{\mu}|$ for all i. Hence, either $\mu = \tilde{\mu}$ or else $y_i = (\mu + \tilde{\mu})/2$ for all i. The latter cannot occur since $y_1, ..., y_p$ are not identical, so it must be the former, i.e., $\phi = \theta$. Second, suppose $|y_i - \mu| \neq |y_i - \tilde{\mu}|$ for all i. By assumption, there exist three distinct elements of $\{y_1, ..., y_p\}$. Without loss of generality, assume these are y_1, y_2, y_3.

Since $\tilde{\beta}_i = \beta_i$ for all i, we have that $|y_i - \mu|/\sqrt{\vartheta} = |y_i - \tilde{\mu}|/\sqrt{\tilde{\vartheta}}$ for $i = 1, 2, 3$. This together with $|y_i - \mu| \neq |y_i - \tilde{\mu}|$ implies $|y_i - \mu| > 0$ and $|y_i - \tilde{\mu}| > 0$, so that

$$\frac{|y_i - \tilde{\mu}|}{|y_i - \mu|} = \frac{\sqrt{\tilde{\vartheta}}}{\sqrt{\vartheta}} = \frac{|y_3 - \tilde{\mu}|}{|y_3 - \mu|}, \quad i = 1, 2.$$

Let $s := (y_3 - \tilde{\mu})/(y_3 - \mu)$ and $t_i := (y_3 - \tilde{\mu})/(y_i - y_3)$. Since $s \neq 0$ and $s + t_i \neq 0$, the above equations yield

$$|(1 + t_i)/(1 + t_i/s)| = |s|, \quad i = 1, 2.$$

Squaring both sides and using $s \neq 1$ yields $-2t_i = s + 1$ for $i = 1, 2$. This implies $t_1 = t_2$, contradicting $y_1 \neq y_2$. Thus, the second case cannot occur, so $\phi = \theta$ is the only possibility. The above argument shows that the converse also holds, i.e., if $\{y_1, ..., y_p\}$ has less than three distinct elements, then we can find $\phi \neq \theta$ for which $c_i(\phi) = c_i(\theta)$ for all i and hence $D(\phi, \theta) = 0$. Thus, the assumption of three distinct elements is necessary and sufficient.

We conclude from the above analysis and Thm. 1 and Lemma 1 that if (i) $\{y_1, ..., y_p\}$ has at least three distinct elements and (ii) $\bar{p}/2 < \alpha(p - \bar{p})$, where $\bar{p}$ is the largest number of identical elements of $\{y_1, ..., y_p\}$, then the sequence $\{\phi^k\}$ generated by the EM algorithm (3) is defined, bounded, and every cluster point is a stationary point of J and thus a stationary point of ℓ.

Notice that if indeed $y_1, ..., y_p$ are generated as assumed, then with probability 1 they are distinct and $\bar{p} = 1$ so that $1/2 < \alpha(p - 1)$ ensures convergence of the EM algorithm. In practice, however, the data may not be generated as assumed and may be subject to measurement errors. Also, if the conditions (i) and (ii) are close to being violated, then the convergence might be slow. To illustrate the conditions (i) and (ii), suppose $p = 3, \beta = 1$ and either (a) $y_1 = 0, y_2 = 0, y_3 = 5$, $\alpha = \frac{1}{4}$ or (b) $y_1 = -5, y_2 = 0, y_3 = 5$, $\alpha = \frac{1}{4}$ or (c) $y_1 = -5, y_2 = 0, y_3 = 5$, $\alpha = \frac{1}{2}$. Then, it can be seen that, in case (a), both (i) and (ii) fail; in case (b), (i) holds but (ii) fails; in case (c) both (i) and (ii) hold. In Figure 1, the graph of J is plotted in each of these three cases. Notice that J is not a concave function.

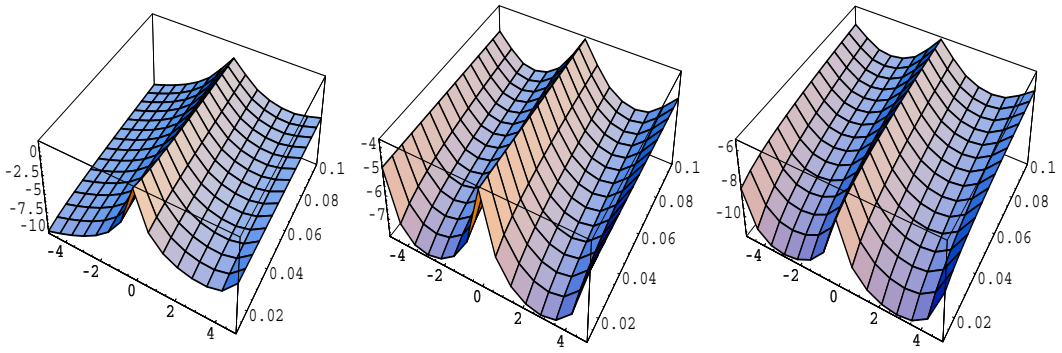

Figure 1: The graph of $J(\mu, \vartheta)$ for the three cases (a), (b), (c). Here μ ranges from -4 to 4 and ϑ ranges from 0^+ to 0.1.

Example 2 A well-known application of the EM algorithm is the estimation a mixture of two Gaussians [9, page 15], [17, Chap. 2]. Here, the observed data $y_1, ..., y_p \in \Re$ are

assumed to be an i.i.d. random sample from a Gaussian density with variance 1 and mean μ_1 if $x_i = 0$ and mean μ_2 if $x_i = 1$. The unobserved variables $x_1, ..., x_p$ form an i.i.d. random sample from the density

$$x_i = \begin{cases} 0 & \text{with probability } \varpi \\ 1 & \text{with probability } 1 - \varpi \end{cases} .$$

The parameter set $\phi = (\mu_1, \mu_2, \varpi)$ is to be estimated. Then, $X = \{0, 1\}$, $\Phi = \Re^2 \times [0, 1]$, and

$$f(x, y|\phi) = \frac{e^{-(y-\mu_1)^2/2}}{\sqrt{2\pi}}\varpi(1 - x) + \frac{e^{-(y-\mu_2)^2/2}}{\sqrt{2\pi}}(1 - \varpi)x.$$

Thus, f does not have the form (12) and so Lemma 3 is not applicable. In what follows, g and ℓ are given by (1), (2), J, ψ, h_i are given by (11), and D is given by (10).

Now, (1), (2), (11) yield

$$\begin{aligned} g(y|\phi) &= \frac{1}{\sqrt{2\pi}}\left(e^{-(y-\mu_1)^2/2}\varpi + e^{-(y-\mu_2)^2/2}(1 - \varpi)\right), \\ J(\phi) &= \sum_{i=1}^{p} \ln\left(e^{-(y_i-\mu_1)^2/2}\varpi + e^{-(y_i-\mu_2)^2/2}(1 - \varpi)\right) - \frac{1}{2}\ln(2\pi). \end{aligned}$$

Thus ∇J is defined and continuous on Φ (since the quantity inside $\ln(\cdot)$ is positive for $\phi \in \Phi$). Moreover, some algebra yields

$$\begin{aligned} D(\phi, \theta) &= \sum_{i=1}^{p} \sum_{x=0,1} \psi\left(\frac{h_i(x|\phi)}{h_i(x|\theta)}\right) h_i(x|\theta), \\ h_i(0|\phi) &= \varpi / \left(\varpi + e^{(y_i-\mu_1)^2/2-(y_i-\mu_2)^2/2}(1 - \varpi)\right), \quad h_i(1|\phi) = 1 - h_i(0|\phi), \ \forall i. \end{aligned}$$

Also, provided that $J(\phi^0) > \max_{\phi\in\Phi:\varpi\in\{0,1\}} J(\phi) = J(\phi_1) = J(\phi_2)$, where $\phi_1 := (\frac{1}{p}\sum_i y_i, 0, 1)$ and $\phi_2 := (0, \frac{1}{p}\sum_i y_i, 0)$, we have that ϖ is uniformly bounded away from 0 and 1 for all $\phi = (\mu_1, \mu_2, \varpi) \in \Phi^0$. Then $h_i(x|\phi)$ is uniformly bounded away from 0 for $x = 0, 1$ and $\phi \in \Phi^0$, implying D and $\nabla_1 D$ are defined and continuous on $\Phi^0 \times \Phi^0$.

The application of the EM algorithm to this example is well known. In particular, the maximization in (3) has a strictly concave objective function and its unique solution is

$$\mu_1^{k+1} = \frac{\sum_{i=1}^p y_i\alpha_i^k}{\sum_{i=1}^p \alpha_i^k}, \quad \mu_2^{k+1} = \frac{\sum_{i=1}^p y_i(1 - \alpha_i^k)}{\sum_{i=1}^p (1 - \alpha_i^k)}, \quad \varpi^{k+1} = \frac{1}{p}\sum_{i=1}^{p} \alpha_i^k, \tag{15}$$

where $\alpha_i^k := h_i(0|\phi^k)$. Clearly, $\phi^{k+1} = (\mu_1^{k+1}, \mu_2^{k+1}, \varpi^{k+1})$ is defined. Moreover, ϕ^{k+1} satisfies (8) with $\delta_k = 0$.

Notice from (15) that if $\phi^0 = \phi_1$, then $\phi^k = \phi_1$ for all $k = 0, 1, ...$, *regardless* of the data. Moreover, direct calculation shows that $\nabla J(\phi_1) \neq 0$ except in special cases such as $y_i = 0$ for all i. The same holds if $\phi^0 = \phi_2$. Thus, we should not choose $\phi^0 = \phi_1$ or $\phi^0 = \phi_2$. Then, unless ϕ_1 and ϕ_2 are maxima for J, we can find (by random sampling or other means) $\phi^0 \in \Phi$ such that $J(\phi^0) > J(\phi_1) = J(\phi_2)$. It is readily shown using induction on k that $0 \leq \varpi^k \leq 1$ and $0 \leq \alpha_i^k \leq 1$, $i = 1, ..., p$, for all k. Then μ_1^{k+1} and μ_2^{k+1}, being

convex combinations of $y_1, ..., y_p$, lie inside the interval $[\min_i y_i, \max_i y_i]$. Hence $\{\phi^k\}$ is bounded. Moreover, $\phi^k \in \Phi^0$ implies ϖ^k is uniformly bounded away from 0 and 1.

Lastly, we verify that $\{\phi^{k+1} - \phi^k\} \to 0$. It can be seen that, by changing μ_1 and μ_2 by the same amount and suitably adjusting ϖ, the value of $h_i(x|\phi)$ would be unchanged (even though its gradient with respect to ϕ changes) for all i and all x. Thus, A3 fails to hold, so Lemma 1 cannot be used as in Example 1. Instead, we verify this directly below. Since $g(y|\phi)$ is bounded above (in fact, $g(y|\phi) \leq 1/\sqrt{2\pi}$), then $J(\phi)$ is bounded above. Since $\{J(\phi^k)\}$ is nondecreasing, this sequence must converge. Since (8) holds with $\delta_k = 0$, this implies $\{D(\phi^{k+1}, \phi^k)\} \to 0$. Since ϖ^k is uniformly bounded away from 0 and 1, then $h_i(x|\phi^k)$ is uniformly bounded away from 0 for $x = 0, 1$ and $i = 1, ..., p$. Since ψ is nonnegative-valued, then the form of D and the fact $\{D(\phi^{k+1}, \phi^k)\} \to 0$ yield

$$\left\{\psi\left(h_i(x|\phi^{k+1})/h_i(x|\phi^k)\right)\right\} \to 0$$

for $x = 0, 1$ and $i = 1, ..., p$. Since $\psi(t) \to 0$ only when $t \to 1$, this in turn implies $\{h_i(x|\phi^{k+1})/h_i(x|\phi^k)\} \to 1$. Since $h_i(x|\phi)$ lies in a bounded interval, this implies $\{h_i(x|\phi^{k+1}) - h_i(x|\phi^k)\} \to 0$. Thus $\{\alpha_i^{k+1} - \alpha_i^k\} \to 0$ for $i = 1, ..., p$. Then $\varpi^{k+1} - \varpi^k = \frac{1}{p}\sum_i(\alpha_i^{k+1} - \alpha_i^k) \to 0$. Also,

$$\mu_1^{k+1} - \mu_1^k = \frac{1}{p}\left(\frac{1}{\varpi^{k+1}} - \frac{1}{\varpi^k}\right)\sum_i y_i\alpha_i^k + \frac{1}{p\varpi^k}\sum_i y_i(\alpha_i^k - \alpha_i^{k-1}).$$

Since $\{\alpha_i^k - \alpha_i^{k-1}\} \to 0$, $\{\varpi^{k+1} - \varpi^k\} \to 0$ and ϖ^k is uniformly bounded away from 0, this shows that $\{\mu_1^{k+1} - \mu_1^k\} \to 0$. A symmetric argument yields $\{\mu_2^{k+1} - \mu_2^k\} \to 0$. Thus $\{\phi^{k+1} - \phi^k\} \to 0$.

We conclude from the above analysis and Thm. 1 that if $J(\phi^0) > J(\phi_1)$, then the sequence $\{\phi^k\}$ generated by the EM algorithm (3) (or, equivalently, (15)) is defined, bounded, and every cluster point is a stationary point of J and thus a stationary point of ℓ. Here, convergence depends crucially on the initial parameter value ϕ^0. If we choose a bad initial value, say $\phi^0 = \phi_1$, then ϕ^k would remain trapped at ϕ_1 which typically is not a stationary point of J nor of ℓ!

6 Further Results on Convergence

The convergence result given in Thm. 1 implies that each cluster point x^∞ of ϕ^k satisfies $\nabla J(\phi^\infty) = 0$. In general, however, a local/global maximum of J over Φ may be a boundary point of Φ where ∇J is not the zero vector and $\nabla_1 D$ is undefined. Can convergence still be shown in this case? This is not easy to answer, especially since J need not be concave.

In this section, we consider the special case of (6) where

B1: $\Phi = \Re_+^m$,

B2: J is defined and continuously differentiable on some open subset of $\Re^m$ containing Φ.

Then $\phi \in \Phi$ is a stationary point of this problem if and only if

$$\nabla_i J(\phi) \leq 0 \text{ and } \nabla_i J(\phi) \cdot \phi_i = 0 \quad i = 1, ..., m, \tag{16}$$

Here $\nabla_i J$ and ϕ_i denote the ith component of ∇J and ϕ, respectively. It is known that any local optimal solution of (6) must be a stationary point [4, pages 195, 196].

For the inexact proximal method (8), instead of (10), we use the corresponding ϕ-divergence for nonnegativity constraints [7, 8, 12, 26, 27], namely,

$$D(\phi, \theta) \ := \ \sum_{i=1}^{m} \psi \left(\frac{\phi_i}{\theta_i} \right) \theta_i, \tag{17}$$

with $\psi \in \Psi$. Notice that $\nabla_1 D$ is not defined on the boundary of Φ, which complicates the analyis of the proximal point method. If J is concave, (6) has an optimal solution, and ψ has some additional properties, then it is known that the sequence $\{\phi^k\}$ generated by (7), (17) is defined and converges to an optimal solution of (6). The proof of this is highly nontrivial–see [8, 12, 27]. We will focus on the case where J is not concave, which requires a different convergence analysis.

The theorem below gives our main convergence result for the proximal point method (8), (17). The result allows for nonconcave J and for optimal solutions lying on the boundary of $\Re^m_+$. However, the result is applicable only for dimensions of $m = 1, 2$ and, even in these cases, the proof is non-trivial.

Theorem 2 *Assume, in addition to B1 and B2, that $\phi^0 \in \Re^m_{++}$ and $\Phi^0 := \{\phi \in \Phi : J(\phi) \geq J(\phi^0)\}$ is bounded. Then $\{\phi^k\}$ generated by (8), with D given by (17) and $\psi \in \Psi$ and $\delta_k \to 0$, has the following properties.*

(a) *ϕ^k is defined and $\phi^k \in \Phi^0 \cap \Re^m_{++}$ for all k.*

(b) *$J(\phi^{k+1}) \geq J(\phi^k)$ for all k, and $\{\nabla_i J(\phi^{k+1}) - \psi' \left(\phi_i^{k+1} / \phi_i^k \right)\} \to 0$, $i = 1, ..., m$.*

(c) *If $m = 1$, then every cluster point of $\{\phi^k\}$ is a stationary point.*

(d) *If $m = 2$, then there exists a cluster point of $\{\phi^k\}$ that is a stationary point.*

Proof. The proof of (a) is by induction on k. Clearly the claim is true for $k = 0$. Suppose the claim is true for some $k \geq 0$. Define $J^k(\cdot) := J(\phi) - D(\cdot, \phi^k)$. If $J^k(\phi) \geq J^k(\phi^k)$, then $J(\phi) \geq J(\phi^k) + D(\phi, \phi^k) \geq J(\phi^k)$. Since $\phi^k \in \Phi^0$, this implies $\phi \in \Phi^0$ and hence

$$\sup_{\phi \in \Phi} J^k(\phi) = \sup_{\phi \in \Phi^0} J^k(\phi) < \infty,$$

where the inequality follows from the boundedness of Φ^0. Also, since ψ is convex, then $\psi(0) := \lim_{t \to 0^+} \psi(t)$ is defined and is either finite or ∞. If $\psi(0) = \infty$, then $\lim_{\phi \to \bar{\phi}} J^k(\phi) = -\infty$ for every $\bar{\phi} \in \Re^m_+ \setminus \Re^m_{++}$. If $\psi(0)$ is finite, then $J^k(\bar{\phi})$ is defined and finite for every $\bar{\phi} \in \Re^m_+ \setminus \Re^m_{++}$. Also, $\lim_{t \to 0^+} \psi'(t) = -\infty$ and $\phi^0 > 0$ imply

$$\lim_{\phi \to \bar{\phi}} \nabla J^k(\phi)^T (\phi^0 - \phi) = \infty,$$

so there exists $\hat{\phi} \in \Re^m_{++}$ such that $J^k(\hat{\phi}) > J^k(\bar{\phi})$. In either case, we see that if $\theta^1, \theta^2, ...$ is a sequence of points in Φ^0 with $J^k(\theta^l) \to \sup_{\phi \in \Phi} J^k(\phi)$ as $l \to \infty$, then any cluster point $\bar{\theta}$ (which exists by boundedness of Φ^0) must be in $\Re^m_{++}$. Since J^k is continuous on

$\Phi^0 \cap \Re^m_{++}$, the maximum is attained at $\bar{\theta}$. Since J^k is continuously differentiable on $\Re^m_{++}$, this implies $\nabla J^k(\bar{\theta}) = 0$. Also, $J^k(\bar{\theta}) \geq J^k(\phi^k)$ implies $J(\bar{\theta}) - D(\bar{\theta}, \phi^k) \geq J(\phi^k)$. Thus, ϕ^{k+1} is defined since taking ϕ^{k+1} to be $\bar{\theta}$ would satisfy (8). Moreover, $\phi^{k+1} \in \Phi^0 \cap \Re^m_{++}$.

(b) follows from (8), (17), $\delta_k \to 0$, and the nonnegativity of D.

We now prove (c). Assume $m = 1$. By (b), $\{J(\phi^k)\}$ is nondecreasing and, by boundedness of Φ^0 and $\phi^k \in \Phi^0$, $\{\phi^k\}$ and $\{J(\phi^k)\}$ are bounded. Let ϕ^∞ be any cluster point of $\{\phi^k\}$. Since J is continuous on Φ^0, then $\lim_{k\to\infty} J(\phi^k) = J(\phi^\infty)$ and, by (8), $\{D(\phi^{k+1}, \phi^k)\} \to 0$. By (17) and $m = 1$,

$$\psi\left(\frac{\phi^{k+1}}{\phi^k}\right)\phi^k \;\to\; 0 \;\text{ as } k \to \infty. \tag{18}$$

Let $\{\phi^k\}_{k\in K}$ ($K \subseteq \{0, 1, ...\}$) be a subsequence that converges to ϕ^∞. Then $\phi^\infty \in \Phi^0$.

If $\phi^\infty > 0$, then (18) implies that, as $k \in K$ and $k \to \infty$,

$$\begin{aligned}
\psi\left(\frac{\phi^{k+1}}{\phi^k}\right) \to 0 \;&\Longrightarrow\; \frac{\phi^{k+1}}{\phi^k} \to 1 \\
&\Longrightarrow\; \psi'\left(\frac{\phi^{k+1}}{\phi^k}\right) \to 0 \text{ and } \phi^{k+1} \to \phi^\infty \\
&\Longrightarrow\; \nabla J(\phi^\infty) = 0,
\end{aligned}$$

where the last implication uses (b).

If $\phi^\infty = 0$, then $\phi^0 > 0$ implies there exist an infinite number of indices ℓ for which $\phi^{\ell-1} > \phi^\ell$. Then, for all $k \in K$ sufficiently large, the index

$$\ell_k := \max\{\ell \in \{1, 2, ..., k\} : \phi^{\ell-1} > \phi^\ell\}$$

is defined and $\ell_k \to \infty$ as $k \to \infty$. Then, $\phi^{\ell_k} \leq \phi^{\ell_k+1} \leq ... \leq \phi^k$, so $\{\phi^{\ell_k}\}_{k\in K} \to 0$. Also,

$$\nabla J(\phi^{\ell_k}) \;<\; \nabla J(\phi^{\ell_k}) - \psi'\left(\frac{\phi^{\ell_k}}{\phi^{\ell_k-1}}\right),$$

where the inequality uses $\phi^{\ell_k}/\phi^{\ell_k-1} < 1$ and $\psi'(t) < 0$ for $t < 1$. Passing to the limit as $k \in K$, $k \to \infty$ and using (b) yields $\nabla J(0) \leq 0$.

We now prove (d). Assume $m = 2$. Arguing as in the proof of (c), we have that $\{\phi^k\}$ is bounded, $\{J(\phi^k)\}$ converges, and $\{D(\phi^{k+1}, \phi^k)\} \to 0$. Since, by (17), D is the sum of $m = 2$ nonnegative terms, then each of these terms tend to 0, i.e.,

$$\psi\left(\frac{\phi_i^{k+1}}{\phi_i^k}\right)\phi_i^k \;\to\; 0 \;\text{ as } k \to \infty, \quad i = 1, 2. \tag{19}$$

For any cluster point $\bar{\phi}$ of $\{\phi^k\}$ and any $i \in \{1, 2\}$, if $\bar{\phi}_i > 0$, then an argument analogous to that used for (c) shows that $\nabla_i J(\bar{\phi}) = 0$. This fact will be used repeatedly below.

Let ϕ^∞ be a cluster point of $\{\phi^k\}$. Let $\{\phi^k\}_{k\in K}$ ($K \subseteq \{0, 1, ...\}$) be a subsequence that converges to ϕ^∞. We consider the following three separate cases.

Suppose $\phi_i^\infty > 0$ for $i = 1, 2$. Then we have that $\nabla_i J(\phi^\infty) = 0$ for $i = 1, 2$ so that ϕ^∞ satisfies (16) and the proof is complete.

Suppose $\phi_1^\infty = 0$ and $\phi_2^\infty > 0$. [The case of $\phi_1^\infty > 0$ and $\phi_2^\infty = 0$ can be treated by swapping the indices 1 and 2.] Then, by defining the index

$$\ell_k := \max\{\ell \in \{1, 2, ..., k\} : \phi_1^{\ell-1} > \phi_1^\ell\}, \tag{20}$$

we obtain, analogous to the argument for (c), that ℓ_k is defined for all k large, $\ell_k \to \infty$ as $k \to \infty$, and any cluster point $\bar{\phi}$ of $\{\phi^{\ell_k}\}_{k \in K}$ (which exists by boundedness of $\{\phi^k\}$) satisfies

$$\bar{\phi}_1 = 0, \quad \nabla_1 J(\bar{\phi}) \leq 0. \tag{21}$$

If $\bar{\phi}_2 > 0$, then we obtain that $\nabla_2 J(\phi^\infty) = 0$ so $\bar{\phi}$ satisfies (16) and the proof is complete. If $\bar{\phi}_2 = 0$, then $\bar{\phi}$ is a cluster point of $\{\phi^k\}$ with $\bar{\phi}_i = 0$ for $i = 1, 2$, so we are in the case below (with $\bar{\phi}$ in place of ϕ^∞).

Suppose $\phi_i^\infty = 0$ for $i = 1, 2$. If $\nabla_i J(\phi^\infty) \leq 0$ for $i = 1, 2$, then ϕ^∞ satisfies (16) and the proof is complete. If $\nabla_i J(\phi^\infty) > 0$ for $i = 1, 2$, then for all ϕ^k in a neighborhood of ϕ^∞ we would have $\nabla_i J(\phi^k) > 0$ for $i = 1, 2$. In this case, (b) and $\psi'(t) > 0$ only if $t > 1$ would imply $\phi_i^{k+1} > \phi_i^k$ for $i = 1, 2$, so ϕ^∞ cannot be a cluster point of $\{\phi^k\}$–a contradiction. Thus, it remains to consider the case of $\nabla_1 J(\phi^\infty) > 0$ and $\nabla_2 J(\phi^\infty) \leq 0$. [The case of $\nabla_1 J(\phi^\infty) \leq 0$ and $\nabla_2 J(\phi^\infty) > 0$ can be treated by swapping the indices 1 and 2.] By defining the index ℓ_k as in (20), we obtain, analogous to the argument for (c), that ℓ_k is defined for all k large, $\ell_k \to \infty$ as $k \to \infty$, and any cluster point $\bar{\phi}$ of $\{\phi^{\ell_k}\}_{k \in K}$ (which exists by boundedness of $\{\phi^k\}$) satisfies (21). Since $\nabla_1 J(\phi^\infty) > 0$, this implies $\bar{\phi} \neq \phi^\infty$. Since $\bar{\phi}_1 = \phi_1^\infty = 0$, this implies $\bar{\phi}_2 \neq \phi_2^\infty = 0$. Since $\bar{\phi}_2 \geq 0$, then $\bar{\phi}_2 > 0$. Since $\bar{\phi}$ is a cluster point of $\{\phi^k\}$, we obtain that $\nabla_2 J(\bar{\phi}) = 0$. This together with (21) implies $\bar{\phi}$ satisfies (16) and the proof is complete. ■

Notice that the above proof for $m = 2$ divides the argument into cases that depend on which face of the orthant $\Re_+^2$ the cluster point lies. We do not know whether the proof can be generalized to $m \geq 3$.

7 Convergence Rate and Convergence Acceleration

It has been observed that the EM algorithm can exhibit slow convergence in practice, as is discussed in [9, 10, 17] and references therein. In this section we give some insights into why this occurs and sketch some schemes to improve the convergence rate.

To motivate our discussion, consider the simple quadratic example:

$$\max_{\phi \in \Re_+} J(\phi) := -\phi^2. \tag{22}$$

The objective function J is strictly concave and the problem has the unique optimal solution $\phi = 0$. If we apply the proximal point method (7) with $\phi^0 > 0$ and D given by (17) and $\psi(t) = -\ln t + t - 1$, we obtain after some calculation that

$$\frac{\phi^{k+1}}{\phi^k} = \frac{2}{\sqrt{1 + 4\phi^k} + 1} < 1$$

for all k. From this it is readily shown that $\{\phi^k\} \to 0$ while $\{\phi^{k+1}/\phi^k\} \to 1$, so the convergence rate is sublinear. The graphical interpretation of this behavior is shown in

Figure 2 (cf. [4, page 546]). We also checked this numerically and, for ϕ^0, it took roughly $k = 500000$ iterations to reach $\phi^k \leq 0.001$, so the convergence is indeed slow. If we perturb the objective function to

$$J(\phi) := -\phi^2 + 0.1\phi,$$

then $\phi = 0$ remains the optimal solution and it took only $k = 61$ iterations to reach $\phi^k \leq 0.001$.

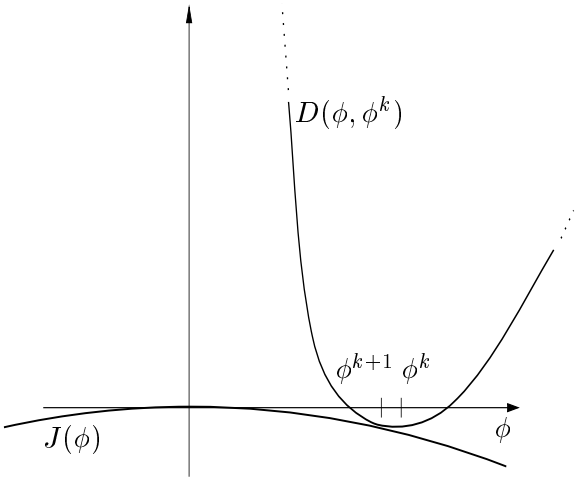

Figure 2: Sublinear convergence with $\{\phi^k\} \to 0$ and $\{\phi^{k+1}/\phi^k\} \to 1$.

What accounts for the slow convergence? It was due to $\nabla J(\phi) = 0$ at the limit point $\phi = 0$. In optimization terminology, strict complementarity is not satisfied at $\phi = 0$. This simple observation suggests that slow convergence of the proximal point method (7), (10) may be due in part to the existence of a cluster point ϕ of $\{\phi^k\}$ and an index $i \in \{1, ..., p\}$ such that $\nabla J(\phi) = 0$ and $h_i(x|\phi) = 0$ for all x in some non-negligible subset of X.

How can the slow convergence be improved? General acceleration schemes for iterative methods such as Aitken's method have been proposed, but these schemes do not exploit the optimization aspect of the problem. Below we sketch some schemes that do exploit this aspect. One scheme would be to regularize the problem by perturbing the objective function J as was done in the above example. However, this can destroy special features of J, such as when J is given by (1), (2), (11) and f has the the exponential form (12). A second scheme would be to construct at iteration k a quadratic function $\tilde{J}$ approximating J if slow convergence is detected. Then a trial point is computed by maximizing $\tilde{J}$ over some "simple" subset $\tilde{\Phi}$ of Φ containing ϕ^k, and this trial point would replace ϕ^k whenever its objective function value is greater than that of ϕ^k. The function $\tilde{J}$ can be constructed, for example, by suitably interpolating through the most recent L iterates and their objective function values: $\phi^k, \phi^{k-1}, ..., \phi^{k-1+L}$ and $J(\phi^k), J(\phi^{k-1}), ..., J(\phi^{k-1+L})$. The simple subset $\tilde{\Phi}$ can be an ellipsoid, for which efficient methods exist to compute the trial point. On the quadratic example (22) with $L = 3$, this scheme would immediately find the optimal solution. A third scheme would be to check whether $h_i(x|\phi^k) \to 0$ for all $i = 1, ..., p$ and all x in some non-negligible subset $\tilde{X}$ of X. If this is occuring, then replace X in the integral defining D with $X \setminus \tilde{X}$ and use other optimization techniques to enforce the

constraints $\int_{\tilde{X}} h_i(x|\phi)dx = 0,\ i = 1, ..., p$. The above description are admittedly sketchy. Their implementation, analysis, and testing are topics for future research.

Acknowledgement. The author thanks Werner Stuetzle for sending me his notes on the EM algorithm, which brought to my attention the Gaussian mixture example in Section 5 and reference [18]. The author also thanks Brad Bell, whose seminar talk on the EM algorithm introduced me to this method.

References

[1] Ash, R. B., *Real Analysis and Probability*, Academic Press, New York, 1972.

[2] Auslender, A., Teboulle, M., and S. A. Ben-Tiba, *Interior proximal and multiplier methods based on second order homogeneous kernels*, Math. Oper. Res., 24 (1999), 645–668.

[3] Ben-Tal, A. and Zibulevsky, M., *Penalty/barrier multiplier methods for convex programming problems*, SIAM J. Optim., 7 (1997), 347–366.

[4] Bertsekas, D. P., *Nonlinear Programming*, 2nd edition, Athena Scientific, Belmont, 1999.

[5] Censor, Y. and Zenios, S. A., *Parallel Optimization: Theory, Algorithms, and Applications*, Oxford University Press, New York and Oxford, 1997.

[6] Chrétien, S. and Hero III, A. O., *Generalized proximal point algorithms*, Report, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, February 2000.

[7] Csiszár, I., *Information-type measures of difference of probability distributions and indirect observations*, Studia Sci. Math. Hungar., 2 (1967), 299–318.

[8] Csiszár, I. and Tusnády, G., *Information geometry and alternating minimization procedures*, Stat. Decisions Suppl., 1 (1984), 205–237.

[9] Dempster, A. P., Laird, N. M., and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Stat. Soc., 39 (1977), 1–38.

[10] Horng, S. C., *Examples of sublinear convergence of the EM algorithm*, in Proceedings of the Statistical Computing Section, American Statistical Association, Alexandria, Virginia; American Statistical Association, 1987, 266–271.

[11] Iusem, A. N. and Teboulle, M., *A regularized dual-based iterative method for a class of image reconstruction problems*, Inverse Problems, 9 (1993), 679–696.

[12] Iusem, A. N. and Teboulle, M., *Convergence rate analysis of nonquadratic proximal methods for convex and linear programming*, Math. Oper. Res., 20 (1995), 657–677.

[13] Kiwiel, K. C., *Proximal minimization methods with generalized Bregman functions*, SIAM J. Control Optim., 35 (1997), 1142–1168.

[14] Luo, Z.-Q. and Tseng, P., *Error bounds and convergence analysis of feasible descent methods: a general approach*, Ann. Oper. Res., 46 (1993), 157–178.

[15] Martinet, B., *Regularisation d'inéquations variationelles par approximations successives*, Rev. Française d'Auto. et Inform. Rech. Opér., 4 (1970),154–159.

[16] Martinet, B., *Determination approchée d'un point fixe d'une application pseudo-contractante*, C. R. Acad. Sci. Paris, 274 (1972), 163–165.

[17] McLachlan, G. J. and Krishnan, T., *The EM Algorithm and Extensions*, Wiley, New York, 1997.

[18] Neal, R. M. and Hinton, G. E., *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in Learning in Graphical Models, edited by M. I. Jordan, Kluwer Academic Publishers, Dordrecht (1998), 355–368; also available from `http://www.cs.toronto.edu/~radford/papers-online.html`

[19] Polyak, R. and Teboulle, M., *Nonlinear rescaling and proximal-like methods in convex optimization*, Math. Programming, 76 (1997), 265–284.

[20] Ortega, J. M. and Rheinboldt, W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[21] Rockafellar, R. T., *Convex Analysis*, Princeton University Press, Princeton, 1970.

[22] Rockafellar, R. T., *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), 877-898.

[23] Rockafellar, R. T., *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), 97–116.

[24] Rockafellar, R. T. and Wets, R.-J.-B., *Variational Analysis*, Springer-Verlag, New York, 1998.

[25] Silva, P. J., Eckstein, J., and C. Humes, Jr., *Rescaling and stepsize selection in proximal method using separable generalized distances*, RUTCOR Research Report 35-99, Rutgers University, Piscataway, 1999.

[26] Teboulle, M., *Entropic proximal mapping with applications to nonlinear programming*, Math. Oper. Res., 17 (1992), 670–690.

[27] Teboulle, M., *Convergence of proximal-like algorithms*, SIAM J. Optim., 7 (1997), 1069–1083.

[28] Tseng, P., *Convergence of block coordinate descent method for nondifferentiable minimization,* J. Optim. Theory Appl., 109 (2001), 473-492.

[29] Tseng, P. and Bertsekas, D. P., *On the convergence of the exponential multiplier method for convex programming*, Math. Programming, 60 (1993), 1–19.

[30] Vardi, Y., Shepp, L. A., and L. Kaufman, *A statistical model for positron emission tomography*, J. Am. Stat. Assoc., 80 (1985), 8–37.

[31] Wu, J. C. F., *On the convergence properties of the EM algorithm*, Annl. Stat., 11 (1983), 95–103.