On EM algorithms and their proximal generalizations

Stéphane Chrétien ∗and Alfred O. Hero †

27 June 2005

Abstract

The goal of this paper is to propose an analysis of the celebrated EM algorithm from the point of view of proximal point algorithms. We define a new type of generalization of the EM procedure called Kullback-proximal algorithms. The proximal framework allows us to prove new results concerning the cluster points. An essential contribution is a detailed analysis of the case where some cluster points lie on the boundary of the parameter space.

keywords: Maximum Likelihood Estimation (MLE), EM algorithm, Proximal Point Algorithm, Karush-Kuhn-Tucker condition, Mixture densities, Competing Risks models.

1 Introduction

The problem of maximum likelihood (ML) estimation consists of finding a solution of the form

\[ \theta_{ML} = \arg\max_{\theta \in \Theta} l_y(\theta), \]

where \( y \) is an observed sample of a random variable \( Y \) defined on a sample space \( \mathcal{Y} \) and \( l_y(\theta) \) is the log-likelihood function defined by

\[ l_y(\theta) = \log g(y; \theta), \]

defined on the parameter space \( \Theta \subset \mathbb{R}^n \), and \( g(y; \theta) \) denotes the density of \( Y \) at \( y \) parametrized by the vector parameter \( \theta \).

The Expectation Maximization (EM) algorithm is an iterative procedure which is widely used for solving ML estimation problems. The EM algorithm was first proposed by Dempster, Laird and Rubin [9] and has seen the number of its potential applications increase substantially since its appearance. The book of McLachlan and Krishnan [14] gives a comprehensive overview of the theoretical properties of the method and its applicability.

The convergence of the sequence of EM iterates towards a maximizer of the likelihood function was claimed in the original paper [9] but it was later noticed that their proof contained a flaw. A careful convergence analysis was finally given by Wu [20] based on Zangwill’s general theory [21]; see also [14]. Zangwill’s theory applies to general iterative schemes and the main task when using it is to verify that the assumptions of Zangwill’s theorems are satisfied. Since the appearance of Wu’s paper, convergence of the EM algorithm is often taken for granted in many cases where the necessary assumptions were sometimes not carefully justified. As an example, an important issue which is often neglected is to understand the behavior of EM iterates when they approach the boundary of the domain of definition of the functions involved. A different example is the following. It is natural to try and establish that EM algorithm actually converges to a point \( \theta^* \), i.e. prove uniqueness of the cluster point. Wu’s approach, reported in [14, Theorem 3.4, p. 89]
is based on the assumption that the euclidean distance between two successive iterates tends to zero. However such an assumption is in fact very hard to verify in most cases and should not be deduced from experimental observations only.

The goal of the present paper is to propose an analysis of EM iterates and their generalizations in the framework of Kullback proximal point algorithms, focusing on the geometric conditions that are indeed provable in practice and underlining the concrete difficulties concerning convergence towards boundaries and cluster point uniqueness. The approach adopted here was first proposed in [5] in which it was shown that the EM algorithm could be recast in the family of Proximal Point algorithms. A proximal scheme for maximizing the function \( l_y(\theta) \) using the distance-like function \( I_y \) is an iterative procedure of the form

\[
\theta^{k+1} \in \arg\max_{\theta \in \Omega} l_y(\theta) - \beta_k I_y(\theta, \theta^k),
\]

where \((\beta_k)_{k \in \mathbb{N}}\) is a sequence of positive real numbers often called relaxation parameters. Proximal point methods were introduced by Martinet [13] and Rockafellar [17] in the context of convex minimization. The proximal point representation proposed in [5] involves the Kullback distance between some well specified conditional densities of a complete data vector and the resulting scheme is thus called Kullback Proximal Point algorithm (KPP). This approach was subsequently further developed in [6] where convergence was studied in the twice differentiable case with the assumption that the limit point be lying in the interior of the domain and where the main novelty was to prove that relaxation of the Kullback-type penalty could ensure superlinear convergence which was confirmed by experimental results on a Poisson linear inverse problem. This paper is an extension of these previous works addressing the main problems of convergence under very general assumptions.

The main results of this paper are the following. Firstly, we prove that all the cluster points of the Kullback proximal sequence which lie in the interior of the domain of definition of the functions involved are stationary points of the likelihood under very mild assumptions that are easily verified in practice. Secondly, taking into account finer properties of \( I_y \), we prove that every cluster point on the boundary of the domain satisfies the Karush-Kuhn-Tucker necessary conditions for optimality under nonnegativity constraints. In order to illustrate our results, we revisit two well known examples: the Poisson inverse problem in medical imaging (Positron Emission Tomography) and the Gaussian mixtures estimation problem. In addition, we apply the Kullback-proximal algorithm to an estimation problem in animal carcinogenicity introduced in [1] in which a very interesting nonconvex constraint has to be handled. In this case, M-steps cannot be obtained in closed form. Still, the Kullback-proximal approach is analyzed and implemented, and numerical experiments are provided which demonstrate the ability of the method to significantly accelerate the convergence of the original EM.

The paper is organized as follows. In Section 2, we recall the basics of the EM algorithm, our Kullback proximal point interpretation and discuss two important examples. Then, Section 3 studies the properties of interior cluster points. We prove that such cluster points are in fact global maximizers of a certain penalized likelihood function. This allows to study the abstract reasons for allowing a relaxation parameter \( \beta \) in the scheme: taking \( \beta \) sufficiently small permits to avoid saddle points. Even more can be proved: if the relaxation sequence is forced to zero, then the cluster points are in fact global maximizers of the likelihood function (at the price of computationally more involved M-steps). Section 4 pursues the analysis in the case where the cluster point lies on some boundary of the domain of \( I_y \). We prove that the Karush-Kuhn-Tucker conditions hold under reasonable assumptions.

2 The Kullback proximal framework

In this section, we review the EM algorithm and the Kullback proximal interpretation of [6].
2.1 The EM algorithm

The EM procedure is an iterative scheme which produces a sequence \((\theta^k)_{k \in \mathbb{N}}\) which is such that each \(\theta^{k+1}\) maximizes a local approximation of the likelihood function in the neighborhood of \(\theta^k\). This point of view will become clear in the proximal point framework of the next subsection.

In the traditional approach, one assumes that some data are hidden from the observer. A frequent example of hidden data is the class to which each sample belongs in the case of mixtures estimation. Another example is when the observed data are projection of a more complete object as for image reconstruction problems in tomography. Thus one would prefer to consider the likelihood of the complete data instead of the ordinary likelihood. Since some parts of the data are hidden, the so called complete likelihood cannot be computed and we therefore have to approximate it. For this purpose, we will need some appropriate notations and assumptions which we now describe.

The observed data are assumed to be i.i.d. samples from a unique random vector \(Y\) taking values on a space \(\mathcal{Y}\). Imagine that we have at our disposal more informative data than just samples from \(Y\). Abstractly, suppose that the more informative data are samples from a random variable \(X\) taking values on a space \(\mathcal{X}\) with density \(f(x; \theta)\) also parametrized by \(\theta\). We will say that the data \(X\) is more informative than the actual data \(Y\) in the sense that \(Y\) is a compression of \(X\), i.e. there exists a non-invertible transformation \(h\) such that \(Y = h(X)\). If one had access to the data \(X\) it would therefore be advantageous to replace the ML estimation problem (1) by

\[
\hat{\theta}_{ML} = \arg\max_{\theta \in \mathbb{R}^d} l_x(\theta),
\]

with \(l_x(\theta) = \log f(x; \theta)\). Since \(y = h(x)\) the density \(g\) of \(Y\) is related to the density \(f\) of \(X\) through

\[
g(y; \theta) = \int_{h^{-1}(\{y\})} f(x; \theta) d\mu(x)
\]

for an appropriate measure \(\mu\) on \(\mathcal{X}\). In this setting, the data \(y\) are called incomplete data whereas the data \(x\) are called complete data.

Of course the complete data \(x\) corresponding to a given observed sample \(y\) are unknown. Therefore, the complete data likelihood function \(l_x(\theta)\) can only be estimated. Given the observed data \(y\) and a previous estimate of \(\theta\) denoted \(\bar{\theta}\), the following minimum mean square error estimator (MMSE) of the quantity \(l_x(\theta)\) is natural

\[
Q(\theta, \bar{\theta}) = \mathbb{E}[\log f(x; \theta) | y; \bar{\theta}],
\]

where, for any integrable function \(F(x)\) on \(\mathcal{X}\), we have defined the conditional expectation

\[
\mathbb{E}[F(x) | y; \bar{\theta}] = \int_{h^{-1}(\{y\})} F(x) k(x | y; \bar{\theta}) d\mu(x)
\]

and \(k(x | y; \bar{\theta})\) is the conditional density function given \(y\)

\[
k(x | y; \bar{\theta}) = \frac{f(x; \bar{\theta})}{g(y; \bar{\theta})}.
\]

Having described the notions of complete data and complete likelihood and its local estimation we now turn to the EM algorithm. The idea is relatively simple : a legitimate way to proceed is to require that iterate \(\theta^{k+1}\) be a maximizer of the local estimator of the complete likelihood conditionally on \(y\) and \(\theta^k\). Hence, the EM algorithm generates a sequence of approximations to the solution (3) starting from an initial guess \(\theta^0\) of \(\hat{\theta}_{ML}\) and is defined by

**Compute** \(Q(\theta, \theta^k) = \mathbb{E}[\log f(x; \theta) | y; \theta^k]\) \hspace{1cm} **E Step**

\[
\theta^{k+1} = \arg\max_{\theta \in \mathbb{R}^d} Q(\theta, \theta^k)
\]

**M Step**

Perhaps the most unsatisfactory point in this standard approach to EM algorithms is that the claimed intention is the one of maximizing an approximate complete likelihood, whereas the actual properties of the EM sequence \((\theta^k)\) are much more modest. We believe that the proximal point approach below provides a less ambiguous motivation for EM procedures and their possible generalizations.
2.2 Kullback proximal interpretation of the EM algorithm

Consider the general problem of maximizing a concave function $\Phi(\theta)$. The original proximal point algorithm introduced by Martinet [13] is an iterative procedure which can be written

$$\theta^{k+1} = \arg\max_{\theta \in D} \left\{ \Phi(\theta) - \frac{\beta_k}{2} \|\theta - \theta^k\|^2 \right\}. \quad (6)$$

The quadratic penalty $\frac{1}{2}\|\theta - \theta^k\|^2$ is relaxed using a sequence of positive parameters $\{\beta_k\}$. In [17], Rockafellar showed that superlinear convergence of this method is obtained when the sequence $\{\beta_k\}$ converges towards zero.

It was proved in [6] that the EM algorithm is a particular example in the class of proximal point algorithms using Kullback Leibler types of penalties. One proceeds as follows. Assume that the family of conditional densities $\{k(x|y;\theta)\}_{\theta \in \mathbb{R}^p}$ is regular in the sense of Ibragimov and Khasminskii [10], in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ are mutually absolutely continuous for any $\theta$ and $\bar{\theta}$ in $\mathbb{R}^p$. Then the Radon-Nikodym derivative $\frac{k(x|y;\theta)}{k(x|y;\bar{\theta})}$ exists for all $\theta, \bar{\theta}$ and we can define the following Kullback Leibler divergence:

$$I_y(\theta, \bar{\theta}) = E\left[\log \frac{k(x|y;\theta)}{k(x|y;\bar{\theta})} \mid y, \bar{\theta}\right]. \quad (7)$$

We are now able to define the Kullback-proximal algorithm. For this purpose, let us define $D_l$ as the domain of $l_y$, $D_{I,\theta}$ the domain of $I_y(\cdot, \theta)$ and $D_I$ the domain of $I_y(\cdot, \cdot)$.

**Definition 2.2.1.** Let $(\beta_k)_{k \in \mathbb{N}}$ be a sequence of positive real numbers. Then, the Kullback-proximal algorithm is defined by

$$\theta^{k+1} = \arg\max_{\theta \in D_l \cap D_{I,\theta}} l_y(\theta) - \beta_k I_y(\theta, \theta^k). \quad (8)$$

The main result on which the present paper relies is the fact that the EM algorithm is a special instance of the Kullback-proximal point algorithm. This was first proved in [6] and then used in the context of mixtures density estimation in [4].

**Proposition 2.2.2.** [6, Proposition 1] The EM algorithm is a special instance of the Kullback-proximal algorithm with $\beta_k = 1$, for all $k \in \mathbb{N}$.

2.3 Examples

Before starting the analysis of Kullback-proximal algorithms, we first present two standard applications of the EM methodology and express the associated M-steps as proximal point iterations.

2.3.1 Poisson inverse problems in Positron Emission Tomography

The model was first proposed and analyzed by Shepp and Vardi [18] and Lange and Carson [12]. Positrons are emitted in a region of an observed body and are received and counted at $d$ detectors. Space is discretized into $n$ points. We assume that emissions occur at point $i$ following a Poisson distribution with parameter $\theta_i$. Conditionally on being emitted at point $i$, a photon is received by detector $j$ with probability $p_{ij}$. This leads to a Poisson model for the counts: if $Y_j$ is the random observed number of photons at detector $j$, then

$$Y_j \sim \mathcal{P}(\mu_j)$$

with

$$\mu_j = \sum_{i=1}^n \theta_i p_{ij}.$$ 

An obvious choice for the complete data vector is $x = (y, z)$ where the vector $z$ consists of the unobservable data counts $z_{ij}$ are the number of photons emitted at point $i$ and detected at
We thus have \( y_j = \sum_{i=1}^{n} z_{ij} \) for \( j = 1, \ldots, d \). It is assumed that given the intensities \( \lambda = (\theta_1, \ldots, \theta_n) \), the associated random counts \( (Z_{ij}) \) are conditionally independent with each \( Z_{ij} \) having a Poisson distribution specified as
\[
Z_{ij} \sim P(\theta_i p_{ij}).
\]
The log-likelihood is given by
\[
l_y(\theta) = \sum_{j=1}^{d} \left( -\sum_{i=1}^{n} \theta_i p_{ij} + y_j \log \left( \sum_{i=1}^{n} \theta_i p_{ij} \right) - \log(y_j!) \right).
\]
The local minimum mean squares estimation of the likelihood is well known to be given by
\[
Q(\theta, \bar{\theta}) = \sum_{j=1}^{d} \sum_{i=1}^{n} \frac{\bar{\theta}_i p_{ij} y_j}{\sum_{l=1}^{n} \bar{\theta}_l p_{lj}} \log(\theta_i p_{ij}) - \theta_i p_{ij}.
\]
Thus, we have
\[
l(\theta) - I_y(\theta, \bar{\theta}) = Q(\bar{\theta}, \bar{\theta}) + K(\bar{\theta}),
\]
where \( K(\bar{\theta}) \) does not depend on \( \theta \). Therefore, we deduce that
\[
I_y(\theta, \bar{\theta}) = Q(\bar{\theta}, \bar{\theta}) - Q(\theta, \bar{\theta}) + \left( l(\theta) - l(\bar{\theta}) \right),
\]
since we a priori require that \( I_y(\theta, \theta) = 0 \). Using this formula, we finally obtain
\[
I_y(\theta, \bar{\theta}) = \sum_{j=1}^{d} \sum_{i=1}^{n} y_j \frac{\bar{\theta}_i}{\sum_{l=1}^{n} \bar{\theta}_l p_{lj}} \log \left( \frac{\theta_i / \sum_{l=1}^{n} \bar{\theta}_l p_{lj}}{\bar{\theta}_i / \sum_{l=1}^{n} \theta_l p_{lj}} \right).
\]
Now, if we adopt the notation
\[
t_{ij}(\theta) = \frac{\theta_i}{\sum_{l=1}^{n} \theta_l p_{lj}},
\]
the distance-like function \( I_y \) becomes
\[
I_y(\theta, \bar{\theta}) = \sum_{j=1}^{d} \sum_{i=1}^{n} y_j t_{ij}(\theta) \log \left( \frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)} \right).
\]
The crucial feature to be noticed in this formula is that the distance-like function \( I_y \) is expressed as the Kullback Leibler divergence between the normalized vectors \( t_{ij}(\theta) \) and \( t_{ij}(\bar{\theta}) \), each coordinate being weighted by the value of \( y_j \).

### 2.3.2 Gaussian mixtures models

We now turn to another estimation problem which is very important in applications, namely the problem of estimating the parameters of a mixture of densities on \( \mathbb{R}^d \). For the purpose of clarity, we restrict the analysis to the Gaussian mixture case, i.e. of a density function of the form
\[
g(y; \theta) = \sum_{j=1}^{J} p_j \frac{1}{(2\pi)^{n/2} \det \Sigma_j} \exp \left( -\frac{1}{2} (y - \mu_j)'^{\Sigma_j^{-1}} (y - \mu_j) \right)
\]
where the \( p_j \)'s (resp. the \( \mu_j \)'s and the \( \Sigma_j \)'s) denote the proportions (resp. the expectations and the covariance matrices) of each gaussian component and
\[
\theta = (p_1, \ldots, p_J, \mu_1, \ldots, \mu_J, \Sigma_1, \ldots, \Sigma_J)
\]
is the parameter vector to be estimated. The parameter space is isomorphic to \( \mathbb{R}^J \times \mathbb{R}^{J \times d} \times \mathbb{R}^{J \times d^2} \) and we denote by \( \Theta \) the affine submanifold which incorporates the constraint that the sum of all proportions equals one, i.e.

\[
\Theta = \left\{ \theta \in \mathbb{R}^J \times \mathbb{R}^{J \times d} \times \mathbb{R}^{J \times d^2} \mid \sum_{j=1}^J \theta_{ij} = 1 \right\}.
\]

The log-likelihood is given by

\[
l_y(\theta) = \sum_{i=1}^n \log \left( \sum_{j=1}^J p_j \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma_j}} \exp \left( -\frac{1}{2} (y_i - \mu_j)^t \Sigma_j^{-1} (y_i - \mu_j) \right) \right).
\]

The complete data are constructed as follows. To each observed data \( y_i \), we associate the label \( z_i \) which takes the value \( j \) if \( y_i \) was drawn from Gaussian component \( j \). Of course, those data are not known to us. The complete data are given by \( x = (y, z) \) and the observed data \( y \) are obtained by projection of \( x \) onto its first \( d \) components, a many to one mapping. The complete density, denoted as above by \( f(x; \theta) \) gives the following expression of the associated complete log-likelihood function \( l_x(\theta) \)

\[
l_x(\theta) = \sum_{i=1}^n \log p_{z_i} + \log \left( \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma_{z_i}}} \exp \left( -\frac{1}{2} (y_i - \mu_{z_i})^t \Sigma_{z_i}^{-1} (y_i - \mu_{z_i}) \right) \right).
\]

The conditional density function \( k(x \mid y; \theta) = f(x; \theta) / g(y; \theta) \) is usually decomposed as follows

\[
k(x \mid y; \theta) = \prod_{i=1}^n t_{iz}(\theta)
\]

where \( t_{ij}(\theta) \) is the conditional probability that \( y_i \) was drawn from the \( j \)th mixture component given \( y \). From Bayes' formula we obtain

\[
t_{ij}(\theta) = \frac{p_j \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma_j}} \exp \left( -\frac{1}{2} (y_i - \mu_j)^t \Sigma_j^{-1} (y_i - \mu_j) \right)}{\sum_{i=1}^J p_i \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma_i}} \exp \left( -\frac{1}{2} (y_i - \mu_i)^t \Sigma_i^{-1} (y_i - \mu_i) \right)}.
\]

Therefore, the Kullback distance-like function \( I_y(\theta, \bar{\theta}) \) has the same form as for the Poisson case, i.e.

\[
I_y(\theta, \bar{\theta}) = \sum_{i=1}^n \sum_{j=1}^J t_{ij}(\bar{\theta}) \log \left( \frac{t_{ij}(\bar{\theta})}{t_{ij}(\theta)} \right).
\]

### 2.4 Notations and assumptions

The notation \( \| \cdot \| \) will be used to denote the norm on any previously defined space without more precision. The space on which it is the norm should be obvious from the context. For any bivariate function \( \Phi, \nabla_1 \Phi \) will denote the gradient with respect to the first variable. In the remainder of this paper we will make the following assumptions.

**Assumptions 2.4.1.** (i) \( l_y \) is differentiable on \( D_l \) and \( l_y(\theta) \) tends to \(-\infty\) whenever \( \| \theta \| \) tends to \(+\infty\).

(ii) the projection of \( D_l \) onto the first coordinate is a subset of \( D_l \).

(iii) \( (\beta_k)_{k \in \mathbb{N}} \) is a convergent nonnegative sequence of real numbers whose limit is denoted by \( \beta^* \).

We will also impose the following assumptions on the distance-like function \( I_y \).
**Assumptions 2.4.2.** (i) There exists a finite dimensional euclidean space $S$, a differentiable mapping $t : D_t \mapsto S$ and a functional $\Psi : D_\Psi \subset S \times S \mapsto \mathbb{R}$ such that

$$I_y(\theta, \bar{\theta}) = \Psi(t(\theta), t(\bar{\theta})),$$

where $D_\Psi$ denotes the domain of $\Psi$.

(ii) For any \{$(t^k, t^k)$\} $\subset D_\Psi$ there exists $\rho_k > 0$ such that $\lim_{k \to \infty} I_y(t^k, t) = \geq \rho_k$. Moreover, we assume that $\inf_{t \in M} \rho_t > 0$ for any bounded set $M \subset S$.

For all $(t', t) \in D_\Psi$, we will also require that

(iii) (Positivity) $\Psi(t', t) \geq 0$,

(iv) (Identifiability) $\Psi(t', t) = 0 \iff t = t'$,

(v) (Continuity) $\Psi$ is continuous at $(t', t)$

and for all $t$ belonging to the projection of $D_\Psi$ onto its second coordinate,

(vi) (Differentiability) the function $\Psi(\cdot, t)$ is differentiable at $t$.

Assumptions 2.4.1(i) and (ii) on $I_y$ are standard and are easily checked in practical examples, e.g. they are satisfied for the Poisson and mixture models considered above. Notice that the domain $D_t$ is now implicitly defined by the knowledge of $D_l$ and $D_\Psi$. Moreover $I_y$ is continuous on $D_l$. The importance of requiring that $I_y$ has the prescribed shape comes from the fact that $I_y$ might not satisfy assumption 2.4.2(iv) in general, as easily checked in both examples above. Therefore assumption 2.4.2 (iv) reflects the need that $I_y$ should at least satisfy the identifiability property up to a non-necessarily injective transformation. In both examples discussed above, this property is an easy consequence of the well known fact that $a \log(a/b) = 0$ implies $a = b$ for positive reals $a$ and $b$. The growth, continuity and differentiability properties 2.4.2 (ii), (v) and (vi) are in any case nonrestrictive.

For the sake of notational convenience, the regularized objective function with relaxation parameter $\beta$ will be denoted

$$F_\beta(\theta, \bar{\theta}) = l_y(\theta) - \beta I_y(\theta, \bar{\theta}). \quad (10)$$

Finally we make the following general assumption.

**Assumptions 2.4.3.** The Kullback proximal iteration (8) is well defined, i.e. there exists at least one maximizer of $F_\beta(\theta, \theta^k)$ at each iteration $k$.

In the EM case, i.e. $\beta = 1$, this last assumption is equivalent to the computability of $M$-steps. A sufficient condition for this assumption to hold would be for instance that $F_\beta(\theta, \bar{\theta})$ be sup-compact, i.e. the level sets $\{\theta \mid F_\beta(\theta, \bar{\theta}) \geq \alpha\}$ be compact for all $\alpha$, $\beta > 0$ and $\theta \in D_t$. However, this assumption is not usually satisfied since the distance-like function is not defined on the boundary of its domain. In practice it suffices to solve the equation $\nabla F_\beta(\theta, \theta^k) = 0$, to prove that the solution is unique. Then, assumption 2.4.1(i) is sufficient to conclude that we actually have a maximizer.

### 2.5 General properties: monotonicity and boundedness

Using Assumptions 2.4.1, we easily deduce monotonicity of the likelihood values and boundedness of the proximal sequence. We start with the following monotonicity result.

**Lemma 2.5.1.** [6, Proposition 2] For any iteration $k \in \mathbb{N}$, the sequence $(\theta^k)_{k \in \mathbb{N}}$ satisfies

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geq \beta k I_y(\theta^k, \theta^{k+1}) \geq 0. \quad (11)$$

We then may deduce the following

**Lemma 2.5.2.** The sequence $(\theta^k)_{k \in \mathbb{N}}$ is bounded.

**Proof.** Due to Lemma 2.5.1, $(\theta^k)_{k \in \mathbb{N}} \subset \{\theta \in \mathbb{R}^n \mid l_y(\theta) \geq l_y(\theta^1)\}$. Then, assumption 2.4.1 (i) on $l_y$ implies the desired result. $\Box$
Lemma 2.5.3. Assume that there exists a subsequence \((\theta^{*})_{k\in\mathbb{N}}\) belonging to a compact set \(C\) included in \(D_{I}\). Then,
\[
\lim_{k \to \infty} \beta_{k}I_{y}(\theta^{k+1}, \theta^{k}) = 0.
\]

Proof. Since \(l_{y}\) is continuous over \(C\), \(\sup_{\theta \in C} l_{y}(\theta) < +\infty\). Thus, \((l_{y}(\theta^{*})_{k\in\mathbb{N}})\) is bounded from above. Moreover, Lemma 2.5.1 implies that the sequence \((l_{y}(\theta^{*}))_{k\in\mathbb{N}}\) is monotone nondecreasing. Therefore, the whole sequence \((l_{y}(\theta^{k}))_{k\in\mathbb{N}}\) is bounded from above and convergent. This implies that \(\lim_{k \to \infty} l_{y}(\theta^{k+1}) - l_{y}(\theta^{k}) = 0\). Applying Lemma 2.5.1 again, we obtain the desired result.

3 Analysis of interior cluster points

The convergence analysis of Kullback proximal algorithms is split into two parts, the first part being the subject of this section. We prove that the accumulation points \(\theta^{*}\) of the Kullback proximal sequence such that \((\theta^{*}, \theta^{*})\) belongs to \(D_{I}\) are stationary points of the log-likelihood function \(l_{y}\).

3.1 Nondegeneracy of the Kullback penalization

We start with the following useful lemma.

Lemma 3.1.1. Let \((\alpha_{1}^{k})_{k\in\mathbb{N}}\) and \((\alpha_{2}^{k})_{k\in\mathbb{N}}\) be two bounded sequences in \(D_{\Psi}\) satisfying
\[
\lim_{k \to \infty} \Psi(\alpha_{1}^{k}, \alpha_{2}^{k}) = 0.
\]
Assume that every couple \((\alpha_{1}^{*}, \alpha_{2}^{*})\) of accumulation points of these two sequences lies in \(D_{\Psi}\). Then,
\[
\lim_{k \to \infty} \|\alpha_{1}^{k} - \alpha_{2}^{k}\| = 0.
\]

Proof. First, one easily obtains that \((\alpha_{2}^{k})_{k\in\mathbb{N}}\) is bounded (use a contradiction argument and Assumption 2.4.2 (ii)). Assume that there exists a subsequence \((\alpha_{1}^{\sigma(k)})_{k\in\mathbb{N}}\) such that \(\|\alpha_{1}^{\sigma(k)} - \alpha_{2}^{\sigma(k)}\| \geq 3\epsilon\) for some \(\epsilon > 0\) and for all large \(k\). Since \((\alpha_{1}^{\sigma(k)})_{k\in\mathbb{N}}\) is bounded, one can extract a convergent subsequence, and thus, we may assume without loss of generality that \((\alpha_{1}^{\sigma(k)})_{k\in\mathbb{N}}\) is convergent with limit \(\alpha^{*}\). Using the triangle inequality, we have \(\|\alpha_{1}^{\sigma(k)} - \alpha_{2}^{\sigma(k)}\| = \epsilon\geq 3\epsilon\). Since \((\alpha_{1}^{\sigma(k)})_{k\in\mathbb{N}}\) converges to \(\alpha_{1}^{*}\), there exists an integer \(K\) such that \(k \geq K\) implies \(\|\alpha_{1}^{\sigma(k)} - \alpha_{2}^{*}\| \leq \epsilon\).

Thus for \(k \geq K\) we have \(\|\alpha_{1}^{k} - \alpha_{2}^{k}\| \geq 2\epsilon\). Now recall that \((\alpha_{2}^{k})_{k\in\mathbb{N}}\) is bounded and extract a convergent subsequence \((\alpha_{2}^{\gamma(k)})_{k\geq K}\) with limit denoted by \(\alpha_{2}^{*}\). Then, using the same arguments as above, we obtain \(\|\alpha_{1}^{*} - \alpha_{2}^{*}\| \geq \epsilon\). Finally, recall that \(\lim_{k \to \infty} \Psi(\alpha_{1}^{k}, \alpha_{2}^{k}) = 0\). We thus have \(\lim_{k \to \infty} \Psi(\alpha_{1}^{\gamma(k)}, \alpha_{2}^{\gamma(k)}) = 0\), and, due to the fact that the sequences are bounded and \(\Psi(\cdot, \cdot)\) is continuous in both variables, we have \(I_{y}(\alpha_{1}^{*}, \alpha_{2}^{*}) = 0\). Thus assumption 2.4.2 (iv) implies that \(\|\alpha_{1}^{*} - \alpha_{2}^{*}\| = 0\) and we obtain a contradiction. Hence, \(\lim_{k \to \infty} \|\alpha_{1}^{k} - \alpha_{2}^{k}\| = 0\) as claimed.

3.2 Cluster points

The main results of this section are the following. First, we prove that under the assumptions 2.4.1, 2.4.2 and 2.4.3, any cluster point \(\theta^{*}\) is a global maximizer of \(F_{\beta^{k}}(\theta^{*}, \theta^{*})\). We then use this general result to prove the natural result that such cluster points are stationary points of the log-likelihood function. We then extend this result to give a natural assumption under which \(\theta^{*}\) is in fact a local maximizer of \(l_{y}\). In addition we show that if the sequence \((\beta^{k})_{k\in\mathbb{N}}\) converges to zero, i.e. \(\beta^{*} = 0\), then \(\theta^{*}\) is a global maximizer of log-likelihood. Finally, we discuss some simple conditions under which the algorithm converges, i.e. has only one cluster point.

The following theorem states a result which describes the stationary points of the proximal point algorithm as global maximizers of the asymptotic penalized function.
**Theorem 3.2.1.** Assume that $\beta^* > 0$. Let $\theta^*$ be any accumulation point of $(\theta^k)_{k \in \mathbb{N}}$. Assume that $(\theta^*, \theta^*) \in D_1$. Then, $\theta^*$ is a global maximizer of the penalized function $F_{\beta^*}(\cdot, \theta^*)$ over the projection of $D_1$ onto its first coordinate, i.e.

$$F_{\beta^*}(\theta^*, \theta^*) \geq F(\theta, \theta^*)$$

for all $\theta$ such that $(\theta, \theta^*) \in D_1$.

An informal argument is as follows. Assume that $\Theta = \mathbb{R}^n$. From the definition of the proximal iterations, we have

$$F_{\beta_{r(k)}}(\theta^{(k)+1}, \theta^{(k)}) \geq F_{\beta_{r(k)}}(\theta, \theta^{(k)})$$

for all subsequence $(\theta^{(k)})_{k \in \mathbb{N}}$ converging to $\theta^*$ and for all $\theta \in \Theta$. Now, assume we can prove that $\theta^{(k)}$ also converges to $\theta^*$, we obtain by taking the limit and using continuity, that

$$F_{\beta_{r}}(\theta^*, \theta^*) \geq F_{\beta_{r}}(\theta, \theta^*)$$

which is the required result. There are two major difficulties when one tries to transform this sketch into a rigorous argument. The first one resides in handling the fact that $l_y$ and $I_y$ are only defined on domains which may not to be closed. Secondly, proving that $\theta^{(k)}$ converges to $\theta^*$ is not an easy task! This issue will be discussed in more details in the next section. The following proof overcomes both difficulties.

**Proof.** Without loss of generality, we may reduce the analysis to the case where $\beta_{k} \geq \beta > 0$ for a certain $\beta$. The fact that $\theta^*$ is a cluster point implies that there is a subsequence of $(\theta^k)_{k \in \mathbb{N}}$ converging to $\theta^*$. For $k$ sufficiently large, we may assume that the terms $(\theta^{(k)+1}, \theta^{(k)})$ belong to a compact neighborhood $C^*$ of $(\theta^*, \theta^*)$ included in $D_1$. Recall that

$$F_{\beta_{r(k)}}(\theta^{(k)}, \theta^{(k)-1}) \geq F_{\beta_{r(k)-1}}(\theta, \theta^{(k)-1})$$

for all $\theta$ such that $(\theta, \theta^{(k)-1}) \in D_1$ and a fortiori for $(\theta, \theta^{(k)-1}) \in C^*$. Therefore,

$$F_{\beta^*}(\theta^{(k)}, \theta^{(k)-1}) - (\beta_k - \beta^*)I_y(\theta^{(k)}, \theta^{(k)-1}) \geq F_{\beta^*}(\theta, \theta^{(k)-1}) - (\beta_{r(k)-1} - \beta^*)I_y(\theta, \theta^{(k)-1}).$$

(12)

Let us have a precise look at the "long term" behavior of $I_y$. First, since $\beta_k > \beta$, for all $k$ sufficiently large, Lemma 2.5.3 says that

$$\lim_{k \to \infty} I_y(\theta^{(k)}, \theta^{(k)-1}) = 0.$$

Thus, for any $\epsilon > 0$, there exists an integer $K_1$ such that $I_y(\theta^{(k)}, \theta^{(k)+1}) \leq \epsilon$ for all $k \geq K_1$.

Moreover, Lemma 3.1.1 and continuity of $t$ allows to conclude that

$$\lim_{k \to \infty} t(\theta^{(k)-1}) = t(\theta^*).$$

Since $\Psi$ is continuous, for all $\epsilon > 0$ and for all $k$ sufficiently large we have

$$I_y(\theta^*, \theta^*) = \Psi(t(\theta^*), t(\theta^*)) \geq \Psi(t(\theta^{(k)}), t(\theta^{(k)-1})) - \epsilon$$

$$= I_y(\theta^{(k)}, \theta^{(k)-1}) - \epsilon.$$  

(13)

On the other hand, $F_{\beta^*}$ is continuous in both variables on $C^*$, due to Assumptions 2.4.1(i) and 2.4.2(i). By continuity in the first and second arguments of $F_{\beta^*}(\cdot, \cdot)$, for any $\epsilon > 0$ there exists $K_2 \in \mathbb{N}$ such that for all $k \geq K_2$

$$F_{\beta^*}(\theta^*, \theta) \leq F_{\beta^*}(\theta^{(k)}, \theta) + \epsilon, $$

(14)

9
On the other hand, using equation 13, since \( l_y \) is continuous, we obtain existence of \( K_3 \) such that for all \( k \geq K_3 \)

\[
F_{\beta^*}(\theta^*, \theta^*) \geq F_{\beta^*}(\theta^{(k)}, \theta^{(k+1)}) - 2\epsilon. \tag{15}
\]

Combining equations (14) and (15) with (12), we obtain

\[
F_{\beta^*}(\theta^*, \theta^*) \geq F_{\beta^*}(\theta^*, \theta) - (\beta_k - \beta^*)l_y(\theta^{(k)}, \theta) + (\beta_k - \beta^*)l_y(\theta^{(k+1)}, \theta^{(k+1)}) - 3\epsilon. \tag{16}
\]

Now, since \( \beta^* = \lim_{k \to +\infty} \beta_k \), there exists an integer \( K_4 \) such that \( \beta_k - \beta^* \leq \epsilon \) for all \( k \geq K_4 \). Then for all \( k \geq \max\{K_1, K_2, K_3, K_4\} \), we obtain

\[
F_{\beta^*}(\theta^*, \theta^*) \geq F_{\beta^*}(\theta^*, \theta) - \epsilon l_y(\theta^{(k)}, \theta) - \epsilon^2 - 3\epsilon.
\]

Since \( l_y \) is continuous and \( (\theta^{(k)})_{k \in \mathbb{N}} \) is bounded, there exists a real constant \( K \) such that \( l_y(\theta^{(k)}, \theta) \leq K \), for all \( n \in \mathbb{N} \). Thus, for all \( k \) sufficiently large

\[
F_{\beta^*}(\theta^*, \theta^*) \geq F_{\beta^*}(\theta^*, \theta) - (4\epsilon K + \epsilon^2). \tag{17}
\]

Finally, recall that no assumption was made on \( \theta \), and that \( C^* \) is any compact neighborhood of \( \theta^* \). Thus, using the assumption 2.4.1(i) which asserts that \( l_y(\theta) \) tends to \( -\infty \) is \( ||\theta|| \) tends to \( +\infty \), we may deduce that (17) holds for any \( \theta \) such that \( (\theta, \theta^*) \in D_I \) and, letting \( \epsilon \) tend to zero, we see that \( \theta^* \) maximizes \( F_{\beta^*}(\theta, \theta^*) \) for over all \( \theta \) such that \( (\theta, \theta^*) \) belongs to \( D_I \) as announced.

Using this theorem, we may now deduce the natural result stating that certain accumulation points away from the boundary of the parameter’s space are stationary points of the log-likelihood function as a simple consequence.

**Corollary 3.2.2.** Assume that \( \beta^* > 0 \). Let \( \theta^* \) be any accumulation point of \( (\theta^k)_{k \in \mathbb{N}} \). Assume that \( (\theta^*, \theta^*) \in \text{int} D_I \). Then, if \( l_y \) is differentiable on \( D_I \), \( \theta^* \) is a stationary point of \( l_y(\theta) \). Moreover, if \( l_y \) is concave, then \( \theta^* \) is a global maximizer of \( l_y \).

**Proof.** Since under the required assumptions \( l_y \) is differentiable and \( l_y(\theta^*, \cdot) \) is differentiable at \( \theta^* \), Theorem 3.2.1 states that

\[
0 \in \nabla l_y(\theta^*) + \beta^* \nabla l_y(\theta^*, \theta^*).
\]

Since \( l_y(\cdot, \theta^*) \) is minimum at \( \theta^* \), \( \nabla l_y(\theta^*, \theta^*) = 0 \) and we thus obtain that \( \theta^* \) is a stationary point of \( l_y \). This is well known to imply that \( \theta^* \) is a global maximizer in the case where \( l_y \) is concave. \( \Box \)

Theorem 3.2.1 seems to be much stronger than the previous corollary. The fact that accumulation points of the proximal sequence may not be global maximizers of the likelihood is now easily seen to be a consequence of the fact that the Kullback distance-like function \( l_y \) perturbs the shape of the likelihood function when \( \theta \) is far from \( \theta^* \). This perturbation does not have serious consequence in the concave case. On the other hand, one may wonder whether \( \theta^* \) cannot be proved to be at least a local maximizer instead of a mere stationary point. The answer is given in the following corollary.

**Corollary 3.2.3.** Let \( \theta^* \) be an accumulation point of \( (\theta^k)_{k \in \mathbb{N}} \) such that \( (\theta^*, \theta^*) \in \text{int} D_I \). In addition, assume that \( l_y \) and \( l_y(\cdot, \theta^*) \) are twice differentiable in a neighborhood of \( \theta^* \) and that the Hessian matrix \( \nabla^2 l_y(\theta^*) \) at \( \theta^* \) is not the null matrix. Then, if \( \beta^* \) is sufficiently small, \( \theta^* \) is a local maximizer of \( l_y \) over \( D_I \).

**Proof.** Assume that \( \theta^* \) is not a local maximizer. Since \( \nabla^2 l_y \) is not the null matrix, for \( \beta^* \) sufficiently small, there is a direction \( \delta \) in the tangent space to \( D_I \) for which the function \( f(t) = F_{\beta^*}(\theta^* + t\delta, \theta^*) \) has positive second derivative for \( t \) sufficiently small. This contradicts the fact that \( \theta^* \) is a global maximizer of \( F_{\beta^*}(\cdot, \theta^*) \) and the proof is completed. \( \Box \)

The next theorem pushes this idea to its end and proves global optimality of accumulation points in the case where the relaxation sequence converges to zero.
Theorem 3.2.4. Let \( \theta^* \) be any accumulation point of \( (\theta^k)_{k \in \mathbb{N}} \). Assume that \((\theta^*, \theta^*) \in D_1\). Then, without assuming differentiability of \( l_y \) nor of \( I_y \), if \((\beta^k)_{k \in \mathbb{N}} \) converges to zero, \( \theta^* \) is a global maximizer of \( l_y \) over the projection of \( D_1 \) along the first coordinate.

Proof. Let \( (\theta^\sigma(k))_{k \in \mathbb{N}} \) be a convergent subsequence of \( (\theta^k)_{k \in \mathbb{N}} \) with limit denoted \( \theta^* \). We may assume that for \( k \) sufficiently large, \((\theta^\sigma(k+1), \theta^\sigma(k)) \) belongs to a compact neighborhood \( C^* \) of \( \theta^* \).

By continuity of \( l_y \), for any \( \epsilon > 0 \), there exists \( K \in \mathbb{N} \) such that for all \( k \geq K \),

\[
l_y(\theta^*) \geq l_y(\theta^\sigma(k)) - \epsilon.
\]

On the other hand, the proximal iteration implies that

\[
l_y(\theta^\sigma(k)) - \beta_{\sigma(k)-1} I_y(\theta^\sigma(k) - 1, \theta^\sigma(k)) \geq l_y(\theta) - \beta_{\sigma(k)-1} I_y(\theta^\sigma(k) - 1, \theta),
\]

for all \( \theta \in D_1 \). Fix \( \theta \in D_1 \). Thus, for all \( k \geq K \),

\[
l_y(\theta^*) \geq l_y(\theta) + \beta_{\sigma(k)-1} I_y(\theta^\sigma(k) - 1, \theta^\sigma(k)) - \beta_{\sigma(k)-1} I_y(\theta^\sigma(k) - 1, \theta) - \epsilon.
\]

Since \( I_y \) is a nonnegative function and \((\beta^k)_{k \in \mathbb{N}} \) is a nonnegative sequence, we obtain

\[
l_y(\theta^*) \geq l_y(\theta) - \beta_{\sigma(k)-1} I_y(\theta^\sigma(k) - 1, \theta) - \epsilon.
\]

Recall that \( (\theta^k)_{k \in \mathbb{N}} \) is bounded due to Lemma 2.5.2. Thus, since \( I_y \) is continuous, there exists a constant \( C \) such that \( I_y(\theta^\sigma(k) - 1, \theta) \leq C \) for all \( k \). Therefore, for \( k \) greater than \( K \),

\[
l_y(\theta^*) \geq l_y(\theta) - \beta_{\sigma(k)-1} C - \epsilon.
\]

Passing to the limit, and recalling that \((\beta^k)_{k \in \mathbb{N}} \) tends to zero, we obtain that

\[
l_y(\theta^*) \leq l_y(\theta) - \epsilon.
\]

Using the same argument as in the end of the proof of Theorem 3.2.1, this latter equation holds for any \( \theta \) such that \((\theta, \theta^*) \) belongs to \( D_1 \), which concludes the proof after letting \( \epsilon \) tend to zero. \( \Box \)

3.3 Convergence of the Kullback proximal sequence

One question remains open in the analysis of the previous section: does the sequence generated by the Kullback proximal point converge? In other words: are there multiple cluster points? In Wu’s paper [20], the answer takes the following form. If the euclidean distance between two successive iterates tends to zero, a well known result states that the set of accumulation points is a continuum (see for instance [16, Theorem 28.1]) and therefore, it is connected. Therefore, if the set of stationary points of \( l_y \) is a countable set, the iterates must converge. As an illustration for this kind of approach, we propose the following theorem.

Theorem 3.3.1. Let \( S^* \) denote the set of accumulation points of the sequence \((\theta^k)_{k \in \mathbb{N}} \). Assume that \( \lim_{k \to \infty} \|\theta^{k+1} - \theta^k\| = 0 \) and that \( l_y(\theta) \) is strictly concave in an open neighborhood \( \mathcal{N} \) of an accumulation point \( \theta^* \) of \((\theta^k)_{k \in \mathbb{N}} \) and that \((\theta^*, \theta^*) \) is in \( \text{int}D_1 \). Then, for any relaxation sequence, the sequence \((\theta^k)_{k \in \mathbb{N}} \) converges to a local maximizer of \( l_y(\theta) \).

Proof. We obtained in Corollary 3.2.2 that every accumulation point \( \theta^* \) of \((\theta^k)_{k \in \mathbb{N}} \) in \( \text{int}D_1 \) and such that \((\theta^*, \theta^*) \in \text{int}D_1 \) is a stationary point of \( l_y(\theta) \). Since \( l_y(\theta) \) is strictly concave over \( \mathcal{N} \), the set of stationary points of \( l_y \) belonging to \( \mathcal{N} \) reduces to singleton. Thus \( \theta^* \) is the unique stationary point in \( \mathcal{N} \) of \( l_y \), and a fortiori, the unique accumulation point of \((\theta^k)_{k \in \mathbb{N}} \) belonging to \( \mathcal{N} \). To complete the proof, it remains to show that there is no accumulation point in the exterior of \( \mathcal{N} \). For that purpose, consider an open ball \( B \) of center \( \theta^* \) and radius \( \epsilon \) included in \( \mathcal{N} \). Then, \( x^* \) is the unique accumulation point in \( B \). Moreover, any accumulation point \( \theta^* \), lying in the exterior of \( \mathcal{N} \) must satisfy \( \|\theta^* - \theta^*\| \geq \epsilon \), and we obtain a contradiction with the fact that \( S^* \) is connected.
Thus every accumulation point lies in \( \mathcal{N} \), from which we conclude that \( \theta^* \) is the only accumulation point of \((\theta^k)_{k \in \mathbb{N}} \) or, in other words, that \((\theta^k)_{k \in \mathbb{N}} \) converges towards \( \theta^* \). Finally, notice that the strict concavity of \( l_y(\theta) \) over \( \mathcal{N} \) implies that \( \theta^* \) is a local maximizer and the proof is completed.

Before concluding this section, let us make two general remarks:

- proving \textit{a priori} that the set of stationary points of \( l_y \) is discrete may be a hard task in particular examples.

- in general, it is not known whether \( \lim_{k \to \infty} \| \theta^{k+1} - \theta^k \| = 0 \) holds. In fact, Lemma 3.1.1 could be a first step in this direction. Indeed if we could prove in any application that the mapping \( t \) is injective, then of course, the desired result would follow immediately. However, injectivity of \( t \) does not hold in any either of the examples of section 2.3. Thus we are now able to clearly understand why the assumption that \( \lim_{k \to \infty} \| \theta^{k+1} - \theta^k \| = 0 \) is not easily deduced from general arguments. This problem has been overcome in [4] where it is shown that \( t \) is componentwise injective and thus performing a componentwise EM algorithm is a good alternative to the standard EM.

4 Analysis of cluster points on the boundary

The goal of this section is to extend the previous results to the case where some cluster points lie on the boundary of the region where computation of proximal steps is well defined. Such cluster points have rarely been analyzed in the statistical literature and the strategy developed for the interior case cannot be applied without further study of the Kullback distance-like function. Notice further that entropic-type penalization terms in proximal algorithms have been the subject of an intensive research effort in the mathematical programming community with the goal of handling positivity constraints; see [19] and the references therein for instance. The analysis proposed here applies to the more general Kullback distances like functions \( I_y \) occurring naturally in the EM context. Our goal is to provide a short proof that such cluster points satisfy the well known Karush-Kuhn-Tucker conditions of nonlinear programming which extend the stationarity condition \( \nabla l_y(\theta) = 0 \) to the case where \( \theta \) is subject to constraints.

In the sequel, the distance-like function will be assumed to have the following additional properties.

Assumptions 4.0.2. The Kullback distance-like function \( I_y \) is of the form

\[
I_y(\theta, \bar{\theta}) = \sum_{1 \leq i \leq n, 1 \leq j \leq m} \alpha_{ij}(y_j) t_{ij}(\theta) \phi \left( \frac{t_{ij}(\theta)}{t_{ij}(\bar{\theta})} \right),
\]

where for all \( i \) and \( j \), \( t_{ij} \) is continuously differentiable on its domain of definition, \( \alpha_{ij} \) is a function from \( \mathcal{Y} \) to \( \mathbb{R}_+ \), the set of positive real numbers and the function \( \phi \) is a non negative convex continuously differentiable function defined for positive reals only and such that \( \phi(\tau) = 0 \) if and only if \( \tau = 1 \).

If \( t_{ij}(\theta) = \theta_i \) and \( \alpha_{ij} = 1 \) for all \( i \) and all \( j \), the function \( I_y \) is the well known \( \phi \) divergence defined by Csiszár in [8]. Assumption 4.0.2 is satisfied in both examples described in section 2.3 with the choice \( \phi(\tau) = \tau \log(\tau) \). This class of entropic distance-like function is rather general and we conjecture that a similar form is obtained for most examples where EM is used.

4.1 More properties of the Kullback distance-like function

The main property that will be needed in the sequel is that under Assumption 4.0.2, the function \( I_y \) satisfies the same property as the one given in Lemma 3.1.1 above, even on the boundary of its domain \( D_I \). This is the result of Proposition 4.1.2 below. We begin with one elementary lemma.
Lemma 4.1.1. Under Assumptions 4.0.2, the function $\phi$ is decreasing on $(0, 1)$, is increasing on $(1, +\infty)$ and $\phi(\tau)$ converges to $+\infty$ when $\tau$ converges to $+\infty$. We have $\lim_{k \to +\infty} \phi(\tau^k) = 0$ if and only if $\lim_{k \to +\infty} \tau^k = 1$.

Proof. The first statement is obvious. For the second statement, the "if" part is trivial, so we only prove the "only if" part. First notice that the sequence $(\tau^k)_{k \in \mathbb{N}}$ must be bounded. Indeed, the level set $\{\tau \mid \phi(\tau) \leq \gamma\}$ is bounded for all $\gamma \geq 0$ and contains the sequence $(\tau^k)_{k \geq K}$ for $K$ sufficiently large. Thus, the Bolzano-Weierstrass theorem applies. Let $\tau^*$ be an accumulation point of $(\tau^k)_{k \in \mathbb{N}}$. Since $\phi$ is continuous, we get that $\phi(\tau^*) = 0$ and thus we obtain $\tau^* = 1$. From this, we deduce that the sequence has only one cluster point, which is equal to 1. Therefore, $\lim_{k \to +\infty} \tau^k = 1$.

Using these lemmas, we are now in position to state and prove the main property of $I_y$.

Proposition 4.1.2. The following statements hold:

(i) For any sequence $(\theta^k)_{k \in \mathbb{N}}$ in $\mathbb{R}^+$ and any bounded sequence $(\eta^k)_{k \in \mathbb{N}}$ in $\mathbb{R}^+$, the fact that $\lim_{k \to +\infty} I_y(\theta^k, \eta^k) = 0$ implies $\lim_{k \to +\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$ for all $i, j$ such that $\alpha_{ij} \neq 0$.

(ii) If one coordinate of one of the two sequences $(\theta^k)_{k \in \mathbb{N}}$ and $(\eta^k)_{k \in \mathbb{N}}$ tends to infinity, so does the other's same coordinate.

Proof. Fix $i$ in $\{1, \ldots, n\}$ and $j$ in $\{1, \ldots, m\}$ and assume that $\alpha_{ij} \neq 0$.

(i) We first assume that $(t_{ij}(\eta^k))_{k \in \mathbb{N}}$ is bounded away from zero. Since $\lim_{k \to +\infty} I_y(\theta^k, \eta^k) = 0$, then $\lim_{k \to +\infty} \phi(t_{ij}(\theta^k)/t_{ij}(\eta^k)) = 0$ and Lemma 4.1.1 implies that $\lim_{k \to +\infty} t_{ij}(\theta^k)/t_{ij}(\eta^k) = 1$. Thus, $\lim_{k \to +\infty} (t_{ij}(\theta^k) - t_{ij}(\eta^k))/t_{ij}(\eta^k) = 0$ and since $t$ is continuous, $t_{ij}(\eta^k)$ is bounded, this implies that $\lim_{k \to +\infty} |t_{ij}(\theta^k) - t_{ij}(\eta^k)| = 0$.

Next, consider the case of a subsequence $(t_{ij}(\eta^k))_{k \in \mathbb{N}}$ which tends towards zero. For contradiction, assume the existence of a subsequence $(t_{ij}(\theta^k))_{k \in \mathbb{N}}$ which remains bounded away from zero, i.e. there exists $a > 0$ such that $t_{ij}(\theta^k)_{k \in \mathbb{N}} \geq a$ for $k$ sufficiently large. Thus, for $k$ sufficiently large we get

$$\frac{t_{ij}(\theta^k)}{t_{ij}(\eta^k)} \geq \frac{a}{t_{ij}(\eta^k)} > 1,$$

and due to the fact that $\phi$ is increasing on $(1, +\infty)$, we get

$$\frac{t_{ij}(\eta^k)\phi(t_{ij}(\theta^k)/t_{ij}(\eta^k))}{t_{ij}(\eta^k)\phi(t_{ij}(\eta^k))} \geq t_{ij}(\eta^k)\phi\left(\frac{a}{t_{ij}(\eta^k)}\right).$$

(18)

On the other hand, Lemma 4.1.1 says that for any $b > 1$, $\phi'(b) > 0$. Since $\phi$ is convex, we get

$$\phi(\tau) \geq \phi(b) + \phi'(b)(\tau - b).$$

Take $\tau = a/t_{ij}(\eta^k)$ in this last expression and combine with (18) to obtain

$$t_{ij}(\eta^k)\phi\left(\frac{t_{ij}(\theta^k)}{t_{ij}(\eta^k)}\right) \geq t_{ij}(\eta^k)\phi(\frac{a}{t_{ij}(\eta^k)} - b).$$

Passing to the limit, we obtain

$$0 = \lim_{k \to +\infty} t_{ij}(\eta^k)\phi\left(\frac{t_{ij}(\theta^k)}{t_{ij}(\eta^k)}\right) \geq a\phi'(b) > 0,$$

which gives the required contradiction.

(ii) If $(t_{ij}(\theta^k))_{k \in \mathbb{N}} \to +\infty$ then, we get $(t_{ij}(\eta^k))_{k \in \mathbb{N}} \to +\infty$ as a direct consequence of part (i). Indeed, if $t_{ij}(\theta^k)$ stays bounded, part (i) says that $\lim_{k \to +\infty} |t_{ij}(\eta^k) - t_{ij}(\theta^k)| = 0$, which contradicts divergence of $(t_{ij}(\theta^k))_{k \in \mathbb{N}}$.

Now, consider the case where $(t_{ij}(\eta^k))_{k \in \mathbb{N}} \to +\infty$. Then, a contradiction is easily obtained if we assume that at least a subsequence $(t_{ij}(\theta^k))_{k \in \mathbb{N}}$ stays bounded from above. Indeed, in such a case, we have

$$\lim_{k \to +\infty} \frac{t_{ij}(\theta^k)}{t_{ij}(\eta^k)} = 0,$$

13
and thus, \( \phi(t_{ij}(\theta^k)/t_{ij}(\eta^k)) \geq \gamma \) for some \( \gamma > 0 \) since we know that \( \phi \) is decreasing on \((0,1)\) and \( \phi(1) = 0 \). This implies that

\[
\lim_{k \to +\infty} t_{ij}(\eta^{(k)}) \phi\left(\frac{t_{ij}(\theta^{(k)})}{t_{ij}(\eta^{(k)})}\right) = +\infty,
\]

which is the required contradiction. \( \square \)

### 4.2 Cluster points are KKT points

The main result of this section is the property that a cluster point \( \theta^* \) such that \( (\theta^*, \theta^*) \) lies on the boundary of \( D_I \) satisfies the Karush-Kuhn-Tucker necessary conditions for optimality on the domain of the log-likelihood function. In the context of Assumptions 4.0.2, \( D_I \) is the set

\[
D_I = \{ \theta \in \mathbb{R}^n \mid t_{ij}(\theta) > 0 \ \forall i \in \{1, \ldots, n\} \text{ and } j \in \{1, \ldots, m\}\}.
\]

We then have the following theorem.

**Theorem 4.2.1.** Let \( \theta^* \) be a cluster point of the Kullback-proximal sequence. Assume that all the functions \( t_{ij} \) are differentiable at \( \theta^* \). Let \( \mathcal{I}^* \) be the set of all couples of indices \((i, j)\) such that the constraint \( t_{ij}(\theta) \geq 0 \) is active at \( \theta^* \), i.e. \( t_{ij}(\theta^*) = 0 \). If \( \theta^* \) lies in the interior of \( D_I \), then \( \theta^* \) satisfies the Karush-Kuhn-Tucker necessary conditions for optimality under the constraint of lying on the active part of the boundary of the closure of \( D_I \), i.e. there exists a family of reals \( \lambda_{ij} \), \((i, j) \in \mathcal{I}^* \) such that

\[
\nabla l_y(\theta^*) + \sum_{(i,j) \in \mathcal{I}^*} \lambda_{ij} \nabla t_{ij}(\theta^*) = 0.
\]

**Proof.** Let \( \Phi_{ij}(\theta, \theta) \) denote the bivariate function defined by

\[
\Phi_{ij}(\theta, \theta) = \phi\left(\frac{t_{ij}(\theta)}{t_{ij}(\theta)}\right).
\]

Let \( \{\theta^{(k)}\}_{k \in \mathbb{N}} \) be a convergent subsequence of the proximal sequence with limit equal to \( \theta^* \). The first order optimality condition at iteration \( k \) is given by

\[
\nabla l_y(\theta^{(k)}) + \beta \phi(\sigma_{ij}(y) \nabla t_{ij}(\theta^{(k)}) \phi\left(\frac{t_{ij}(\theta^{(k)})}{t_{ij}(\theta^{(k)})}\right) + \sum_{ij} \alpha_{ij}(y) t_{ij}(\theta^{(k)}) \nabla \Phi(\theta^{(k)}, \theta^{(k)})\right) = 0.
\]

Now, we have

\[
t_{ij}(\theta^{(k)}) \nabla \Phi(\theta^{(k)}, \theta^{(k)}) = -t_{ij}(\theta^{(k)}) \phi'(t_{ij}(\theta^{(k)})) \nabla t_{ij}(\theta^{(k)})
\]

for all \( i \) and \( j \). In fact, this last expression converges to zero as established now.

**Claim A.** For all \((i, j)\) such that \( \alpha_{ij}(y_j) \neq 0 \), we have

\[
\lim_{k \to +\infty} t_{ij}(\theta^{(k)}) \nabla \Phi(\theta^{(k)}, \theta^{(k)}) = 0.
\]

**Proof claim A.** Two cases may occur. In the first case, we have \( t_{ij}(\theta^*) = 0 \). Since the sequence \( \{\theta^k\}_{k \in \mathbb{N}} \) is bounded due to Lemma 2.5.2, countinuous differentiability of \( \phi \) and the \( t_{ij} \) proves that \( \nabla \Phi(\theta^{(k)}, \theta^{(k)}) \) is bounded from above. Thus, the desired conclusion follows. In the second case, \( t_{ij}(\theta^*) \neq 0 \) and applying Lemma 2.5.3, we deduce that \( I_y(\theta^{(k)}, \theta^{(k)}) \) tends to zero. Hence, we get that \( \lim_{k \to +\infty} \Phi(\theta^{(k)}, \theta^{(k)}) = 0 \), which implies that \( \lim_{k \to +\infty} \theta^{(k)} / \theta^{(k)} = 1 \).

From this and Assumptions 4.0.2, we deduce the that \( \lim_{k \to +\infty} \phi'(t_{ij}(\theta^{(k)})/t_{ij}(\theta^{(k)})) = 0 \). Since \( \{\theta^{(k)}\}_{k \in \mathbb{N}} \) converges to \( \theta^* \) and that \( t_{ij}(\theta^*) \neq 0 \), we obtain that the subsequence
we obtain the same result with the additional conclusion that the for every cluster point $\beta$ for the Gaussian mixture problem satisfy Assumptions 2.4.1, 2.4.2, 2.4.3 and 4.0.2. If the family exactly the Karuch-Kuhn-Tucker necessary condition for optimality as claimed.  

$\theta$ that $t$ the components $D$ closure of $5.1$ Back to the Poisson inverse problem and the mixture estimation examples Then, we analyze a nonparametric survival analysis with competing risks proposed by Ahn, on the meaning of our results in the context of the two well known examples described in Section 2.3. The goal of this section is to present some applications of the previous theory. We first comment 5 Applications Proposition 5.1.1. The Kullback proximal algorithm for the Poisson linear inverse problem and for the Gaussian mixture problem satisfy Assumptions 2.4.1, 2.4.2, 2.4.3 and 4.0.2.  

Proof. This can be readily checked with the precaution of taking $t = (t_{ij})$ but not including the components $t_{ij}$ where $j$ is such that $y_j = 0$. □

\[ \{t_{ij}(\theta^{(k)-1})/t_{ij}(\theta^{(k)})\}_{k \in \mathbb{N}} \] is bounded from above. Moreover, $\{\nabla t_{ij}(\theta^{(k)})\}_{k \in \mathbb{N}}$ is also bounded by count 'limous differentiability of $t_{ij}$. Therefore, the fact that $\lim_{k \to +\infty} \phi'(t_{ij}(\theta^{(k)-1})/t_{ij}(\theta^{(k)})) = 0$ allows to conclude the proof of the claim.

Using this claim, we just have to study the remaining right hand side terms in (19), namely the expression $\sum_{ij} \alpha_{ij}(y_j) \nabla t_{ij}(\theta^{(k)}) \phi \left( \frac{t_{ij}(\theta^{(k)-1})}{t_{ij}(\theta^{(k)})} \right)$. Let $\mathcal{I}^{**}$ be a subset of the active indices $\mathcal{I}$ such that the family $\nabla t_{ij}(\theta^*)$ is linearly independent. This linear independence is preserved under small perturbations, we may assume without loss of generality that the family $\left(\nabla t_{ij}(\theta^{(k)})\right)_{(i,j) \in \mathcal{I}^{**}}$ is linearly independent for $k$ sufficiently large. For such $k$, we may rewrite equation (19) as

\[
\nabla l_y(\theta^{(k)}) + \beta_\sigma (k) \left( \sum_{(i,j) \in \mathcal{I}^{**}} \lambda^{\sigma(k)}(y_j) \nabla t_{ij}(\theta^{(k)}) + \sum_{ij} \alpha_{ij}(y_j) t_{ij}(\theta^{(k)}) \nabla \Phi(\theta^{(k)}, \theta^{(k)-1}) \right) = 0. \tag{20}
\]

We prove the following claim.

Claim B. The sequence $\{\lambda^{\sigma(k)}(y_j)\}_{k \in \mathbb{N}}$ is bounded.

Proof of Claim B. Using the previous claim and the continuous differentiability of $l_y$ and the $t_{ij}$, equation (20) expresses that the $\lambda^{\sigma(k)}(y_j)$ are proportional to the coordinates of the projection on the span of the $\nabla t_{ij}(\theta^{(k)})$ of a vector converging towards $\nabla l_y(\theta^*)$. Since the $\nabla t_{ij}(\theta^{(k)})$, for $(i,j) \in \mathcal{I}^{**}$, form a linearly independent family for $k$ sufficiently large, none of the coordinates can tend towards infinity.

We are now in position of finishing the proof of the theorem. Take any cluster point $\tau_{ij}$ of $t_{ij}(\theta^{(k)-1})/t_{ij}(\theta^{(k)})$. Using the claim above, we know that $\left(\lambda^{\sigma(k)}_{ij}(y_j)\right)_{(i,j) \in \mathcal{I}^{**}}$ lies in a compact set. Let $\left(\lambda^{\sigma(k)}_{ij}\right)_{(i,j) \in \mathcal{I}^{**}}$ be a cluster point of this sequence. Passing to the limit, we obtain from equation (19) that

\[
\nabla l_y(\theta^{(k)}) + \beta^* \left( \sum_{(i,j) \in \mathcal{I}^{**}} \lambda^*_j \nabla t_{ij}(\theta^*) \right) = 0.
\]

for every cluster point $\beta^*$ of $\{\beta(\theta^{(k)})\}_{k \in \mathbb{N}}$. For all $(i,j) \in \mathcal{I}^{**}$, set $\lambda_{ij} = \beta^* \lambda^*_j$. This equation is exactly the Karuch-Kuhn-Tucker necessary condition for optimality as claimed. □

Remark 4.2.2. If the family $\left(\nabla t_{ij}(\theta^{(k)})\right)_{(i,j) \in \mathcal{I}^{**}}$ is linearly independent for $k$ sufficiently large, we obtain the same result with the additional conclusion that the $\lambda_{ij}$ are nonnegative, which proves that $\theta^*$ satisfies the Karush-Kuhn-Tucker conditions for the constraints of simply lying in the closure of $D_1$. 

5 Applications

The goal of this section is to present some applications of the previous theory. We first comment on the meaning of our results in the context of the two well known examples described in Section 2.3. Then, we analyze a nonparametric survival analysis with competing risks proposed by Ahn, Kodell and Moon in [1].

5.1 Back to the Poisson inverse problem and the mixture estimation examples

We begin with the following result.

Proposition 5.1.1. The Kullback proximal algorithm for the Poisson linear inverse problem and for the Gaussian mixture problem satisfy Assumptions 2.4.1, 2.4.2, 2.4.3 and 4.0.2. 

Proof. This can be readily checked with the precaution of taking $t = (t_{ij})$ but not including the components $t_{ij}$ where $j$ is such that $y_j = 0$. □
It is important to see how Theorem 4.2.1 applies in the context of our favorite examples. In the case of the Poisson inverse problem described in Section 2.3.1, the expression of the log-likelihood forbids that all components of \( \{ \theta_n \}_{n \in \mathbb{N}} \) tend to zero. It seems natural in practice to assume that the conditional probabilities \( p_{ij} \) are such that \( t_{ij}(\theta) = 0 \) if and only if \( \theta_t = 0 \). It is possible for the Kullback proximal iterates to approach the situation where \( t_{ij} \) vanishes for some indices \( i \) and \( j \), corresponding to the physical situation where some pixels \( \theta_i \) are likely not to have emitted a positron. Therefore, in such cases, the corresponding cluster points have to satisfy the Karush-Kuhn-Tucker necessary conditions for optimality instead of simply being stationary points of the likelihood.

Let us now turn to the case of Gaussian mixtures estimation of Section 2.3.2. It may happen that a proportion \( p^0_j \) tends to zero along the iterates. In such a case, any cluster point must satisfy the Karush-Kuhn-Tucker involving the terms \( \nabla t_{ij} \) for all \( i \). As before, this type of cluster point may not be simply a stationary point of the log-likelihood function. On the other hand, it is possible to find cluster points \( \theta^* \) lying on the boundary of \( D_I \) without satisfying Karush-Kuhn-Tucker conditions. This case happens if one of the covariance matrices tends to be singular. If so, the log-likelihood tends to \( +\infty \) when approaching this cluster point, which is therefore often called a degenerate solution of the problem. The assumptions of Theorem 4.2.1 exclude such possible situations for which differential optimality conditions are obviously not appropriate: indeed such cluster points do not lie in the interior of \( D_I \) as required. On the other hand, degeneracy in Gaussian mixtures estimation via the EM algorithm is carefully studied in [2].

### 5.2 Attribution of tumor lethality

In this subsection, we apply the proximal viewpoint to the problem of nonparametric estimation in survival analysis with competing risks proposed by Ahn, Kodell and Moon in [1].

#### 5.2.1 Presentation of the problem

This problem can be described as follows. Consider a group of animals in an animal carcinogenicity experiment. Sacrifices are performed at certain prescribed times denoted by \( t_1, t_2, \ldots, t_m \) in order to study the presence of the tumor of interest. Let \( T_1 \) be the time to onset of the tumor of interest, \( T_D \) the time to death from this tumor and \( X_C \) be the time to death from a cause other than this tumor. Among the quantities to be estimated are \( S(t) \), \( P(t) \) and \( Q(t) \), the survival function of \( T_1 \), \( T_D \) and \( X_C \) respectively. It is assumed that \( T_1 \) and \( T_D \) are independent of \( X_C \).

A nonparametric approach is proposed: observed data \( y_1, \ldots, y_n \) are the number of deaths on every interval \( [t_j, t_{j+1}] \) which can be classified into several categories, namely

- death with tumor (without knowing cause of death)
- death without tumor
- sacrifice with tumor
- sacrifice without tumor

This gives rise to a multinomial model whose probability mass is parametrized by the values of \( S \), \( P \) and \( Q \) at times \( t_1, \ldots, t_m \). More precisely, if for each time interval \( [t_j, t_{j+1}] \) we denote by \( c_j \) the number of deaths with present tumor, \( b_{ij} \) the number of deaths without, \( a_{2j} \) the number of sacrifices with tumor and \( b_{2j} \) the number of sacrifices without tumor and if we let \( N_j \) the number of live animals at \( t_j \), it is shown in [1] that the corresponding log-likelihood is given by

\[
\log g(y; \theta) = \sum_{j=1}^m (N_{j-1} - N_j) \sum_{k=1}^{N_j} \log(p_k q_k) + (a_{2j} + b_{2j}) \log(p_j q_j) + c_j \log \left( \frac{(1 - p_j)(1 - \pi_j p_j)(1 - q_j)}{(1 - p_j)(1 - \pi_j p_j)(1 - q_j)} \right) + b_{ij} \log((1 - q_j) \pi_j - 1) + a_{2j} \log(1 - \pi_j) + b_{2j} \log \pi_j + Cst
\]
where \( Cst \) is a constant \( \pi_j = S(t_j)/P(t_j) \), \( p_j = P(t_j)/P(t_{j-1}) \) and \( q_j = Q(t_j)/Q(t_{j-1}) \) for all \( j = 1, \ldots, m \), \( \theta = (\pi_1, \ldots, p_j, p_1, \ldots, p_j, q_1, \ldots, q_j) \) and the parameter space is specified by the constraints

\[
\Theta = \left\{ \theta = (\pi_1, \ldots, p_j, p_1, \ldots, p_j, q_1, \ldots, q_j) \mid 0 \leq \pi_j \leq 1, \\
0 \leq p_j \leq 1, \quad 0 \leq q_j \leq 1, \quad j = 1, \ldots, m \right. \\
\left. \quad \text{and } \pi_j p_j \leq \pi_{j-1} \quad j = 2, \ldots, m \right\},
\]

(22)

where the last constraint serves to impose monotonicity of \( S \) while monotonicity of \( P \) and \( Q \) is a direct consequence of the constraints on the \( p_j \)'s and the \( q_j \)'s respectively.

In order to use the EM or the Kullback proximal methodology, a complete data set needs to be introduced. Natural complete data \( x_1, \ldots, x_n \) should fall into one of the following categories

- death caused by tumor
- death with incidental tumor
- death without tumor
- sacrifice with tumor
- sacrifice without tumor

and the many-to-one mapping \( h \) consists of merging the first two of them. To each time interval \( (t_j, t_{j+1}] \) there correspond the numbers \( d_j \) of deaths caused by tumor and \( a_{1j} \) of deaths with incidental tumor, none of which being observable. The associated complete log-likelihood function is given by

\[
\log f(x; \theta) = \sum_{j=1}^{m} (N_{j-1} - N_j) \sum_{k=1}^{j-1} \log(p_k q_k) + (a_{2j} + b_{2j}) \log(p_j q_j) \\
+ d_j \log(1 - p_j) + a_{1j} \log \left( (1 - \pi_j p_j)(1 - q_j) \right) \\
+ b_{1j} \log((1 - q_j) \pi_{j-1}) + a_{2j} \log(1 - \pi_j) + b_{2j} \log \pi_j + Cst
\]

(23)

Now, we have to compute the expectation \( Q(\theta, \bar{\theta}) \) of the log-likelihood function of the complete data conditionally to the parameter \( \bar{\theta} \). This is done in the following way: the random variables \( d_j \) and \( a_{1j} \) can be assumed binomial with parameter \( \lambda_j \) and \( 1 - \lambda_j \) where \( \lambda_j \) is the probability that the death was caused by the tumor conditionally to the presence of the tumor. Conditionally to \( \bar{\theta} \), we have

\[
\lambda_j = \frac{1 - \bar{p}_j}{1 - \bar{p}_j + (1 - \pi_j p_j)(1 - q_j)}
\]

(24)

(see [1, Section 3] for details). From this, we obtain that the conditional mean values of \( d_j \) and \( a_{1j} \) are given by

\[
E[d_j \mid y; \bar{\theta}] = \lambda_j c_j \quad \text{and} \quad E[a_{1j} \mid y; \bar{\theta}] = (1 - \lambda_j)c_j.
\]

(25)

Thus, we get

\[
Q(\theta, \bar{\theta}) = \sum_{j=1}^{m} (N_{j-1} - N_j) \sum_{k=1}^{j-1} \log(p_k q_k) + (a_{2j} + b_{2j}) \log(p_j q_j) \\
+ \lambda_j c_j \log(1 - p_j) + (1 - \lambda_j)c_j \log \left( (1 - \pi_j p_j)(1 - q_j) \right) \\
+ b_{1j} \log((1 - q_j) \pi_{j-1}) + a_{2j} \log(1 - \pi_j) + b_{2j} \log \pi_j + Cst.
\]

(26)

From this, we can easily compute the associated Kullback distance-like function and obtain

\[
I_p(\theta, \bar{\theta}) = \sum_{j=1}^{m} c_j \left( t_j(\theta) \phi \left( \frac{t_j(\bar{\theta})}{t_j(\theta)} \right) + t'_j(\theta) \phi \left( \frac{t'_j(\bar{\theta})}{t'_j(\theta)} \right) \right)
\]

(27)

with

\[
t_j(\theta) = \frac{1 - p_j}{1 - p_j + (1 - \pi_j p_j)(1 - q_j)} \quad \text{and} \quad t'_j(\theta) = \frac{(1 - \pi_j p_j)(1 - q_j)}{1 - p_j + (1 - \pi_j p_j)(1 - q_j)}
\]

(28)
and as in our previous examples $\phi$ defined by $\phi(\tau) = \tau \log(\tau)$. Hence, in this problem Assumptions 2.4.1, 2.4.2 2.4.3 and 4.0.2 are again satisfied.

The main computational problem in this example is to handle some difficult nonconvex constraints entering the definition of the parameter space $\Theta$. The authors of [1] propose to use the method called the Complex Method proposed by Box in [3]. Using our proximal point framework, we are able to easily incorporate the nonconvex constraints into the Kullback distance-like function. For this purpose, let $I'_y$ be defined by

$$I'_y(\theta, \bar{\theta}) = I_y(\theta, \bar{\theta}) + \sum_{j=1}^{m} t''_j(\theta) \phi\left(\frac{t''_j(\bar{\theta})}{t''_j(\theta)}\right)$$

(29)

where

$$t''_j(\theta) = \pi_j - \pi_j p_j.$$  

(30)

Using this new function, the nonconvex constraints $\pi_j p_j \leq \pi_{j-1}$ are naturally satisfied for all proximal iterations, while Assumptions 4.0.2 are kept unviolated.

5.2.2 Experimental results

We implemented the Kullback proximal algorithm with different choices of constant relaxation sequence $(\beta_k)_{k \in \mathbb{N}}$. Since each iteration cannot be given in closed form, we computed the successive iterates using a rudimentary gradient algorithm with constant stepsize using the Scilab software. Some of our results for the data of Table 3. (MCL, Male, AL) of [1] are given in Figure 1. Our stopping criterion is the following: stop when the increase in the log likelihood is less than $10^{-3}$.

The main conclusions from these experiments are the following. Firstly, the proximal approach for different values of $\beta$ was very easy to implement and since M-steps cannot be obtained in closed form for the standard EM, our proximal scheme did not increase the complexity compared to EM.

Secondly, although we did not try to program a refined implementation of the optimization subroutines for the proximal steps using Scilab, we noticed that convergence occurred in a quite reasonable time (less than thirty seconds on a PC) while the C++ implementation given in [15, Section 5] of Box’s method takes about one minute on a SUN Spark 10 workstation. These computational results cannot however be compared since both algorithms use different stopping criteria, different programming languages and different computers. Nevertheless our method is a real contribution to the computational aspect of this competing risks problem since Box’s method used in [1] and [15] is not proved to converge to a stationary point of the likelihood. See the reviewer’s warning in the Math. Reviews.

Thirdly, the experiments clearly show the role of the relaxation parameter $\beta$ in accelerating the convergence speed of the method. Finally, we did not notice any dependency of the obtained parameter with respect to the initialization, which seems to indicate a certain robustness of our scheme although nonconvex constraints were handled.

Finally, our method

6 Conclusions

The goal of this paper was to study the convergence behavior of the EM algorithm and its proximal generalizations. We clarified greatly the analysis by making use of the Kullback-proximal framework. Two of our main contributions are the following. Firstly we showed that interior cluster points are stationary points of the likelihood function and are local maximizers for sufficiently small values of $\beta$. Secondly, we showed that cluster points lying on the boundary satisfy the Karush-Kuhn-Tucker conditions. Such cases were very seldom studied in the literature although constrained estimation is a topic of growing importance; see for instance the special issue of the Journal of Statistical Planning and Inference [11] which is devoted to the problem of estimation under constraints. Also, on the negative side, the analysis from the Kullback-proximal viewpoint allowed us to understand why uniqueness of the cluster point is hard to establish theoretically.
Figure 1: Evolution of the log-likelihood versus iteration number
On the positive side, we were able to implement a new and efficient proximal point method for estimation in a hard problem involving nonlinear inequality constraints.

References

[1] Ahn, H., Moon, H., Kodell, R. L., ”Attribution of tumour lethality and estimation of the time to onset of occult tumours in the absence of cause-of-death information”. J. Roy. Statist. Soc. Ser. C, vol. 49, no. 2, 157–169, 2000.

[2] Biernacki, C. and Chretien, S., ”Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM”. Statist. Probab. Lett. 61 (2003), no. 4, 373–382.

[3] M.J. Box, ”A new method of constrained optimization and a comparison with other methods”, The Computer Journal, 8, 42–52, 1965.

[4] G. Celeux, S. Chretien, F. Forbes and A. Mkhadri, ”A component-wise EM algorithm for mixtures”, J. Comput. Graph. Statist. 10 (2001), no. 4, 697–712 and INRIA RR-3746, Aug. 1999.

[5] S. Chretien and A.O. Hero, ”Acceleration of the EM algorithm via proximal point iterations”, Proceedings of the International Symposium on Information Theory, MIT, Cambridge, p. 444, 1998.

[6] S. Chretien and A. Hero, ”Kullback proximal algorithms for maximum-likelihood estimation,” IEEE Trans. Inform. Theory 46 (2000), no. 5, 1800–1810.

[7] T. Cover and J. Thomas, Elements of Information Theory, Wiley, New York, 1987.

[8] I. Csiszár, ”Information-type measures of divergence of probability distributions and indirect observations”, Studia Sci. Math. Hung., 2 (1967), 299–318.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, ”Maximum likelihood from incomplete data via the EM algorithm,” J. Royal Statistical Society, Ser. B, vol. 39, no. 1, pp. 1–38, 1977.

[10] I. A. Ibragimov and R. Z. Has’minskii, Statistical estimation: Asymptotic theory, Springer-Verlag, New York, 1981.

[11] Journal of Statistical Planning and Inference, vol. 107, no. 1–2, 2002.

[12] K. Lange and R. Carson, ”EM reconstruction algorithms for emission and transmission tomography,” Journal of Computer Assisted Tomography, vol. 8, no. 2, pp. 306–316, 1984.

[13] B. Martinet, ”Régularisation d’inéquation variationelles par approximations successives,” Revue Française d’Informatique et de Recherche Operationnelle, vol. 3, pp. 154–179, 1970.

[14] McLachlan, G. J., Krishnan, T., ”The EM algorithm and extensions,” Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley and Sons, Inc., New York, 1997.

[15] Moon, H., Ahn, H. Kodell, R. and Pearce, B. ” A comparison of a mixture likelihood method and the EM algorithm for an estimation problem in animal carcinogenicity studies,” Computational Statistics and Data Analysis, 31 , no. 2, pp. 227–238, 1999.

[16] Ostrowski, A. M., Solution of equations and systems of equations. Pure and Applied Mathematics, Vol. IX. Academic Press, New York-London 1966

[17] R. T. Rockafellar, ”Monotone operators and the proximal point algorithm,” SIAM Journal on Control and Optimization, vol. 14, pp. 877–898, 1976.
[18] L. A. Shepp and Y. Vardi, “Maximum likelihood reconstruction for emission tomography,” *IEEE Trans. on Medical Imaging*, vol. MI-1, No. 2, pp. 113–122, Oct. 1982.

[19] M. Teboulle, “Entropic proximal mappings with application to nonlinear programming,” *Mathematics of Operations Research*, vol. 17, pp. 670–690, 1992.

[20] C. F. J. Wu, “On the convergence properties of the EM algorithm,” *Annals of Statistics*, vol. 11, pp. 95–103, 1983.

[21] W. I. Zangwill and B. Mond, *Nonlinear programming: a unified approach*, Prentice-Hall International Series in Management. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969.