Kullback Proximal Point Algorithms for Maximum Likelihood Estimation

Stéphane Chrétien and Alfred O. Hero III *

November 25, 1998

Abstract

In this paper, we study the convergence of a new class of fast and stable sequential optimization methods for computing maximum likelihood estimates. These methods are based on a *proximal point algorithm* implemented with a Kullback-type proximal function. When the proximal regularization parameter is set to unity one obtains the classical expectation maximization (EM) algorithm. For other values of the regularization parameter, relaxed versions of EM are obtained which can have much faster convergence. In particular, if the regularization parameter is made to depend on iteration in a suitable way, a superlinearly converging algorithm is obtained. We present an implementation of the algorithm using the *trust region* update strategy. For illustration the method is applied to a non-quadratic inverse problem with Poisson distributed data.

Keywords 1 *accelerated EM algorithm, Kullback regularization, proximal point iterations, superlinear convergence, trust region methods.*

1 Introduction

Iterative solutions to the maximum likelihood (ML) estimation problem are of interest when direct closed form solution is infeasible. Among the most

*Dept. of Electrical Engineering and Computer Science, The University of Michigan, 1301 Beal Avenue, Ann Arbor, Michigan 48109-2122, USA, Email chretien,hero@eecs.umich.ed

stable iterative strategies to ML is the popular expectation maximization (EM) algorithm [?]. This algorithm has the attractive property of monotonicity which guarantees that the likelihood function increases with each iteration. The convergence properties of the EM algorithm and its variants have been extensively studied in the literature; see [?] and [?] for instance. It is well known that under strong concavity assumptions the EM algorithm converges linearly towards the ML estimator θ_{ML}. However, in practice the EM algorithm suffers from slow convergence in late iterations. Efforts to improve the asymptotic convergence of the EM algorithm include: Aitken's acceleration [?], overrelaxation [?], conjugate gradient [?] [?], Newton methods [?] [?], quasi-Newton methods [?], ordered subsets EM [?] and stochastic EM [?]. Unfortunately, these methods do not automatically garantee the monotone increasing likelihood property of standard EM, which frequently requires additional monitoring for instability [?].

The main goal of this paper is to recast the EM algorithm into a more general framework of monotone algorithms having the potential for accelerated convergence. For this purpose, the EM algorithm is identified as a particular instance of a proximal point algorithm using Kullback regularization. The proximal point algorithm, first introduced by Rockafellar [?] and Martinet [?], is a state of the art procedure in optimization. In particular, proximal approaches have led to many high performance numerical algorithms; e.g. bundle methods for nonsmooth problems [?] and multiplier methods for constrained optimization [?]. A key motivation for the proximal point algorithm is that an iteration-dependent penalty can be introduced to obtain superlinear convergence rates in the case of quadratic regularization [?]. In this paper, this idea is used to obtain relaxed versions of EM algorithm with superlinear asymptotic convergence rates.

The outline of the paper is the following. In Section 2 we provide a brief review of key elements of the classical EM algorithm. In Section 3, we establish a relationship between the EM algorithm and the proximal point algorithm. In section 4, we present the general Kullback proximal point algorithm and we establish global and superlinear convergence to the maximum likelihood estimator. In section 5, we study second order approximations of the Kullback proximal point iteration using trust region updating. The trust region strategy is introduced in order to obtain global convergence and monotonic behavior of the approximate scheme. Finally, in Section **??** we present numerical comparisons for a Poisson inverse problem.

Our notations are standard for the most part. $\nabla_{10} I(\theta, \bar{\theta})$ (resp. $\nabla^2_{10} I(\theta, \bar{\theta})$)

denotes the gradient (resp. the hessian matrix) of $I(\theta, \bar{\theta})$ in the first variable.

2 Background

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

$$\theta_{ML} = \text{argmax}_{\theta \in \mathbb{R}^p} \; l_y(\theta), \tag{1}$$

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

$$l_y(\theta) = \log g(y; \theta), \tag{2}$$

where $g(y;\theta)$ denotes the density of Y at y parametrized by a deterministic vector θ in $\mathbb{R}^p$. One of the most popular methods for solving ML estimation problems is the Expectation Maximization (EM) algorithm described in Dempster, Laird, and Rubin [?] which we describe as follows.

A more informative data space $\mathcal{X}$ is introduced. A random variable X is defined on $\mathcal{X}$ that with density $f(x;\theta)$ parametrized by θ. 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 $Y = h(X)$. It would therefore be advantageous to replace the ML estimation problem (1) by

$$\theta_{ML} = \max_{\theta \in \mathbb{R}^p} l_x(\theta), \tag{3}$$

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

$$g(y;\theta) = \int_{h^{-1}(\{y\})} f(x;\theta) d\mu(x) \tag{4}$$

for an appropriate measure μ on $\mathcal{X}$. Under condition (4) and for a given observed sample y, for any x in $h^{-1}(\{y\})$ the solutions of (3) are solutions of the original ML estimation problem (1). In this setting, the data y are called *incomplete data* whereas the data x are called *complete data*.

It remains to deal with the fact that 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 θ denoted $\bar{\theta}$, the following minimum mean square error estimator (MMSE) of the quantity $l_x(\theta)$ is natural

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

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

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

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

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

The EM algorithm generates a sequence of approximations to the solution (3) starting from an initial guess θ^0 of θ_{ML} and defined by

$$\textbf{Compute } Q(\theta, \theta^k) = \mathsf{E}[\log f(x; \theta)|y; \theta^k] \qquad \textbf{E Step}$$

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} Q(\theta, \theta^k) \qquad \textbf{M Step}$$

The key to understanding the convergence of the EM algorithm is the following decomposition of the log-likelihood function

$$l_y(\theta) = Q(\theta, \bar{\theta}) + H(\theta, \bar{\theta}) \tag{5}$$

where

$$H(\theta, \bar{\theta}) = -\mathsf{E}[k(x|y; \theta)|y; \bar{\theta}].$$

Furthermore, it follows from elementary application of Jensen's inequality to the log function that

$$H(\theta, \bar{\theta}) \geq H(\theta, \theta) \geq 0, \quad \forall \theta, \ \bar{\theta} \in \mathbb{R}^p. \tag{6}$$

Observe from (5) and (6) that the θ function $\{Q(\theta, \theta^k)\}$ is a lower bound on the log likelihood function $l_y(\theta)$. This fact is necessary to ensure monotonicity of the algorithm. Specifically, using the M-step defining relation

$$Q(\theta^{k+1}, \theta^k) \geq Q(\theta^k, \theta^k), \tag{7}$$

one obtains

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geq Q(\theta^{k+1}, \theta^k) - Q(\theta^k, \theta^k) \tag{8}$$

$$+ H(\theta^{k+1}, \theta^k) - H(\theta^k, \theta^k). \tag{9}$$

Hence, using (7) and (6)

$$l_y(\theta^{k+1}) \geq l_y(\theta^k).$$

This is the well known monotonicity property of the EM algorithm.

Note that if the function $H(\theta, \bar{\theta})$ in (5) were scaled by an arbitrary positive factor β the function $Q(\theta, \bar{\theta})$ would remain a lower bound on $l_y(\theta)$ and monotonicity of the algorithm would be preserved. As will be shown below, if β is allowed to vary with iteration in a suitable manner one obtains a monotone, superlinearly convergent generalization of the EM algorithm.

3 Proximal methods and the EM algorithm

In this section, we present the proximal point algorithm of Rockafellar and Martinet. We then demonstrate that EM is a particular instance of a proximal point method implemented with a Kullback-type penalty.

3.1 The proximal point algorithm

Consider the general problem of maximizing a concave function $\Phi(\theta)$. Then, the proximal point algorithm is an iterative procedure which can be written

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \left\{ \Phi(\theta) - \frac{\beta_k}{2} ||\theta - \theta^k||^2 \right\}. \tag{10}$$

The quadratic penalty $||\theta - \theta^k||^2$ is relaxed using a sequence of positive parameters $\{\beta_k\}$. In [?], Rockafellar showed that superlinear convergence of this method is obtained when the sequence $\{\beta_k\}$ converges towards zero. In numerical implementations of proximal point the function $\Phi(\theta)$ is generally replaced by a piecewise linear model [?].

3.2 Proximal interpretation of the EM algorithm

We now turn to the EM algorithm and demonstrate that the EM procedure can be recast into a the proximal point framework. For this purpose, we will need to consider a particular Kullback information measure. Assume that the family $k(x|y;\theta)$ with $\theta \in \mathbb{R}^p$ is regular in the sense of Ibragimov and Khasminskii [?], in particular $k(x|y;\theta)\mu(x)$ and $k(x|y;\bar{\theta})\mu(x)$ are absolutely continuous with respect to each other for any θ and $\bar{\theta}$ in $\mathbb{R}^p$. Then the Radon-Nikodym derivative $\frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}$ exists for all $\theta, \bar{\theta}$ and the following Kullback measure is well defined.

$$I(\theta, \bar{\theta}|y) = \mathsf{E}\big[\log \frac{k(x|y,\bar{\theta})}{k(x|y;\theta)}|y;\bar{\theta}\big]. \tag{11}$$

Proposition 1 *The EM algorithm is a proximal point algorithm with Kullback-type penalty (11) of the form*

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \left\{ l_y(\theta) - I(\theta, \bar{\theta}|y) \right\} \tag{12}$$

Proof 1 *The key to making the connection with the proximal point algorithm is the following representation of the M step:*

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathsf{E}\big[\log \frac{f(x;\theta)}{g(y;\theta)}|y;\theta^k\big]\big\}.$$

This equation is equivalent to

$$\begin{aligned} \theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) &+ \mathsf{E}\big[\log \frac{f(x;\theta)}{g(y;\theta)}|y;\theta^k\big] \\ &- \mathsf{E}\big[\log \frac{f(x;\theta^k)}{g(y;\theta^k)}|y;\theta^k\big]\big\} \end{aligned}$$

since the additional term does not modify the maximization problem. Recalling that $k(x|y;\theta) = \frac{f(x;\theta)}{g(y;\theta)}$,

$$\begin{aligned} \theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) &+ \mathsf{E}\big[k(x|y;\theta)|y;\theta^k\big] \\ &- \mathsf{E}\big[k(x|y;\theta^k)|y;\theta^k\big]. \end{aligned}$$

We finally obtain

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \big\{ \log g(y;\theta) + \mathsf{E}\big[\frac{k(x|y;\theta)}{k(x|y;\theta^k)}|y;\theta^k\big]\big\}$$

which concludes the proof.

4 Kullback Generalization of Proximal Point Algorithm

The proximal point interpretation of EM in the previous section suggests that EM can be generalized by replacing the quadratic penalty in the standard proximal point recursion (10) with the Kullback penalty (11). Other non-quadratic generalizations of proximal point have been proposed by Teboulle and others [**?**]. In this section we introduce the Kullback generalization of proximal point, called the Kullback proximal point (KPP) algorithm, and study its convergence properties.

4.1 Algorithm Definition

Let $k(x;\theta)$ be any family of densities defined on a set $\mathcal{S} \subset \mathcal{X}$, parametrized by θ and regular in the sense of Ibragimov and Khasminskii [**?**]. Define the Kullback information measure

$$I(\theta,\bar{\theta}) = \int_{\mathcal{S}} \log \frac{k(x;\theta)}{k(x;\bar{\theta})} k(x;\bar{\theta}) d\mu(x). \tag{13}$$

Notice that for the choice $k(x;\theta) = k(x|y;\theta)$ and for $S = h^{-1}(\{y\})$, we obtain $I(\theta,\bar{\theta}) = I(\theta,\bar{\theta}|y)$ as defined by (11) in the context of the EM algorithm. We make the following assumptions, where Λ_M denotes the greatest eigenvalue of the matrix M and λ_M denotes the smallest.

Assumptions 1 *We assume the following:*

(i) $l_y(\theta)$ *is twice continuously differentiable on* $\mathbb{R}^p$ *and* $I(\theta,\bar{\theta})$ *is twice continuously differentiable in* θ *for all* $\bar{\theta}$ *in* $\mathbb{R}^p$*.*

(ii) $\lim_{\|\theta\|\to\infty} l_y(\theta) = -\infty$.

(iii) $l_y(\theta) < \infty$ *and* $\lambda_{\nabla^2 l_y(\theta)} \leq \Lambda_{\nabla^2 l_y(\theta)} < 0$ *on every bounded* θ*-set*

(iv) for any $\bar{\theta}$ *in* $\mathbb{R}^p$, $I(\theta,\bar{\theta}) < \infty$ *and* $0 < \lambda_{\nabla^2_{10} I(\theta,\bar{\theta})} \leq \Lambda_{\nabla^2_{10} I(\theta,\bar{\theta})}$ *on every bounded* θ*-set.*

These assumptions ensure smoothness of $l_y(\theta)$ and $I(\theta,\bar{\theta})$ and their first two derivatives in θ. Assumption 1.iii also implies strict concavity of $l_y(\theta)$. Assumption 1.iv implies that $I(\theta,\bar{\theta})$ is strictly convex and that the parameter

θ is strongly identifiable in the family of densities $k(x;\theta)$ (see proof of Lemma 1 below). Note that the above assumptions are not the minimum set possible, e.g. that $l_y(\theta)$ and $I(\theta,\bar{\theta})$ are upper bounded follows from continuity, Assumption 1.ii and the property $I(\theta,\bar{\theta}) \geq I(\bar{\theta},\bar{\theta}) = 0$, respectively.

With the above assumptions we are now prepared to introduce the general algorithm.

Definition 1 *Let $g(y;\theta)$ and $k(x;\theta)$ be such that $l_y(\theta)$ and $I(\theta,\bar{\theta})$ satisfy Assumptions 1. Let $\{\beta_k\}$ be a sequence of positive relaxation parameters. Then, the following recurrence will be called the Kullback proximal point (KPP) algorithm*

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \big\{ l_y(\theta) - \beta_k I(\theta, \theta^k) \big\}. \tag{14}$$

The KPP algorithm is well defined since the maximum in (14) is always achieved in a bounded set. Monotonicity is guaranteed by this procedure as proved in the following proposition.

Proposition 2 *The log-likelihood sequence $\{l_y(\theta^k)\}$ is monotone non-decreasing and satisfies*

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geq \beta_k I(\theta^{k+1}, \theta^k), \tag{15}$$

Proof 2 *From iteration (14), we have*

$$l_y(\theta^{k+1}) - l_y(\theta^k) \geq \beta_k I(\theta^{k+1}, \theta^k) - \beta_k I(\theta^k, \theta^k).$$

Since $I(\theta^k, \theta^k) = 0$ and $I(\theta^{k+1}, \theta^k) \geq 0$, we deduce (15) and that $\{l_y(\theta^k)\}$ is non-decreasing.

We next turn to convergence of the KPP iterates $\{\theta^k\}$.

4.2 Convergence

First we characterize the fixed points of the KPP algorithm.

Proposition 3 *Given $\beta_k = \beta > 0$, $k = 1, 2, \ldots$, the fixed points of the iteration (14) are maximizers of the log-likelihood function $l_y(\theta)$.*

Proof 3 *Consider a fixed point θ^* of iteration (14) for $\beta_k = \beta =$ constant. Then,*

$$\theta^* = \text{argmax}_{\theta \in \mathbb{R}^p} \{l_y(\theta) - \beta I(\theta, \theta^*)\}.$$

As $l_y(\theta)$ and $I(\theta, \theta^*)$ are both smooth in θ, θ^* must be a stationary point

$$0 = \nabla l_y(\theta^*) - \beta \nabla_{10} I(\theta^*, \theta^*).$$

Furthermore, as $I(\theta, \bar{\theta})$ has a minimum (zero) at $\theta = \bar{\theta}$, $\nabla_{10} I(\theta^*, \theta^*) = 0$,

$$0 = \nabla l_y(\theta). \tag{16}$$

Since $l_y(\theta)$ is strictly concave, we deduce that θ^* is a maximizer of $l_y(\theta)$.

Remark 1 *For nonconvex log-likelihood, equation (16) asserts that fixed points of iteration (14) for constant β_k are stationary points of $l_y(\theta)$. Stationary points can be minimizers, maximizers or saddle points.*

The following will be useful.

Lemma 1 *Let $k(x; \theta)$ be such that Assumptions 1 are satisfied for $I(\theta, \bar{\theta})$. Then, given two bounded sequences $\{\theta_1^k\}$ and $\{\theta_2^k\}$, $\lim_{k \to \infty} I(\theta_1^k, \theta_2^k) = 0$ implies that $\lim_{k \to \infty} \|\theta_1^k - \theta_2^k\| = 0$.*

Proof 4 *Let $\mathcal{B}$ be any bounded set containing both sequences $\{\theta_1^k\}$ and $\{\theta_2^k\}$. Let λ denote the minimum*

$$\lambda = \min_{\theta, \bar{\theta} \in \mathcal{B}} \lambda_{\nabla_{10}^2 I(\theta, \bar{\theta})} \tag{17}$$

Assumption 1.iv implies that $\lambda > 0$. Furthermore, invoking Taylor's theorem with remainder, $I(\theta, \bar{\theta})$ is strictly convex in the sense that for any k

$$I(\theta_1^k, \theta_2^k) \geq I(\theta_1^k, \theta_1^k) + \nabla I(\theta_1^k, \theta_1^k)^{\mathsf{T}} (\theta_1^k - \theta_2^k) + \frac{1}{2} \lambda \|\theta_1^k - \theta_2^k\|^2.$$

As $I(\theta_1^k, \theta_1^k) = 0$ and $\nabla_{10} I(\theta_1^k, \theta_1^k) = 0$ we obtain

$$I(\theta_1^k, \theta_2^k) \geq \frac{\lambda}{2} \|\theta_1^k - \theta_2^k\|^2.$$

The desired result comes from passing to the limit $k \to \infty$.

Using these results, we easily obtain the following.

Lemma 2 *Let $g(y;\theta)$ and $k(x;\theta)$ be such that Assumptions 1 are satisfied. Then $\{\theta^k\}$ is bounded.*

Proof 5 *Due to Proposition 2, the sequence $\{l_y(\theta^k)\}$ is monotone increasing. Therefore, assumption 1.ii implies that $\{\theta^k\}$ is bounded.*

In the following lemma, we prove a result which is often called asymptotic regularity [**?**].

Lemma 3 *Let the densities $g(y;\theta)$ and $k(x;\theta)$ be such that Assumptions 1 are satisfied on $l_y(\theta)$ and $I(\theta,\bar{\theta})$. Assume in addition that $\beta_k = \phi(\|\theta^k - \theta^{k-1}\|)$ with $\phi(t) : \mathbb{R}_+ \mapsto \mathbb{R}_+$ satisfying*
a. $\phi(t)$ is positive and continuous for all $t \in \mathbb{R}_+$,
b. $\phi(t) = 0 \Leftrightarrow t = 0$.
Then $\lim_{k\to\infty} \|\theta^{k+1} - \theta^k\| = 0$.

Proof 6 *By Assumption 1.iii and by Proposition 2 $\{l_y(\theta^k)\}$ is bounded and monotone. Since, by Lemma 2, $\{\theta^k\}$ is a bounded sequence $\{l_y(\theta^k)\}$ converges. Therefore, $\lim_{k\to\infty} \{l_y(\theta^{k+1}) - l_y(\theta^k)\} = 0$ which, from (15), implies*

$$\lim_{k\to\infty} \beta_k I(\theta^{k+1}, \theta^k) = 0. \tag{18}$$

Thus, either $\lim_{k\to\infty} \beta_k = 0$ or $\lim_{k\to\infty} I(\theta^{k+1}, \theta^k) = 0$. In the first case, the assumptions on the function $\phi(t)$ give the desired result. In the second case, Lemma 1 establishes the result.

We can now give a global convergence theorem.

Theorem 1 *Under the same assumptions as in Lemma 3, $\{\theta^k\}$ converges to the solution of the ML estimation problem (1).*

Proof 7 *Since $\{\theta^k\}$ is bounded, one can extract a convergent subsequence $\{\theta^{\sigma(k)}\}$ with limit θ^*. The defining iteration (14) implies that*

$$\nabla l_y(\theta^{\sigma(k)+1}) - \beta_{\sigma(k)} \nabla_{10} I(\theta^{\sigma(k)+1}, \theta^k) = 0.$$

Due to Assumptions 1.i, $\nabla l_y(\theta)$ is continuous in θ and $\nabla_{10} I(\theta,\bar{\theta})$ is continuous in $(\theta,\bar{\theta})$. Hence,

$$\nabla l_y(\theta^*) - \phi(0) \nabla_{10} I(\theta^*, \theta^*) = 0$$

or, as the second term on the left of the inequality is zero

$$\nabla l_y(\theta^*) = 0.$$

As, by Assumption 1.iii, $l_y(\theta)$ is concave, θ^* is a maximizer of $l_y(\theta)$ so that θ^* solves the Maximum Likelihood estimation problem (1). Furthermore, as positive definiteness of $\nabla^2 l_y$ implies that $l_y(\theta)$ is in fact strictly concave, this maximizer is unique. Hence, $\{\theta^k\}$ has only one accumulation point and $\{\theta^k\}$ converges to θ^* which ends the proof.

We now establish the main result concerning speed of convergence.

Theorem 2 *Under the same assumptions as in Lemma 3, $\{\theta^k\}$ converges surperlinearly to the solution of the ML estimation problem (1).*

Proof 8 *Due to Theorem 1, the sequence $\{\theta^k\}$ converges to the unique maximizer θ_{ML} of $l_y(\theta)$. Assumption 1.i implies that the gradient mapping $\nabla_\theta\big(l_y(\theta) - \beta_k I(\theta, \theta_{ML})\big)$ is continuously differentiable. Hence, we have the following Taylor expansion about θ_{ML}.*

$$\begin{aligned} \nabla l_y(\theta) - \beta_k \nabla_{10} I(\theta, \theta_{ML}) = {} & \nabla l_y(\theta_{ML}) \\ & - \beta_k \nabla_{10} I(\theta_{ML}, \theta_{ML}) \\ & + \nabla^2 l_y(\theta_{ML})(\theta - \theta_{ML}) \\ & - \beta_k \nabla^2_{10} I(\theta_{ML}, \theta_{ML})(\theta - \theta_{ML}) \\ & + R(\theta - \theta_{ML}), \end{aligned} \tag{19}$$

where the remainder satisfies

$$\lim_{\theta \to \theta_{ML}} \frac{\|R(\theta - \theta_{ML})\|}{\|\theta - \theta_{ML}\|} = 0.$$

Since θ_{ML} maximizes $l_y(\theta)$, $\nabla l_y(\theta_{ML}) = 0$. Furthermore, $\nabla_{10} I(\theta_{ML}, \theta_{ML}) = 0$. Hence, (19) can be simplified to

$$\begin{aligned} \nabla l_y(\theta) - \beta_k \nabla_{10} I(\theta, \theta_{ML}) = {} & \nabla^2 l_y(\theta_{ML})(\theta - \theta_{ML}) \\ & - \beta_k \nabla^2_{10} I(\theta_{ML}, \theta_{ML})(\theta - \theta_{ML}) + R(\theta - \theta_{ML}). \end{aligned} \tag{20}$$

From the defining relation (14) the iterate θ^{k+1} satisfies

$$\nabla l_y(\theta^{k+1}) - \beta_k \nabla_{10} I(\theta^{k+1}, \theta^k) = 0. \tag{21}$$

So, taking $\theta = \theta^{k+1}$ in (20) and using (21), we obtain

$$\begin{aligned}
&\beta_k\big(\nabla_{10}I(\theta^{k+1},\theta^k) - \nabla_{10}I(\theta^{k+1},\theta_{ML})\big) = \\
&+ \nabla^2 l_y(\theta_{ML})(\theta^{k+1}-\theta_{ML}) - \beta_k\nabla^2_{10}I(\theta_{ML},\theta_{ML})(\theta^{k+1}-\theta_{ML}) \\
&+ R(\theta^{k+1}-\theta_{ML}).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\|\beta_k\big(\nabla_{10}I(\theta^{k+1},\theta^k) - \nabla_{10}I(\theta^{k+1},\theta_{ML})\big) - R(\theta^{k+1}-\theta_{ML})\| = \\
&\|\nabla^2 l_y(\theta_{ML})(\theta^{k+1}-\theta_{ML}) - \beta_k\nabla^2_{10}I(\theta_{ML},\theta_{ML})(\theta^{k+1}-\theta_{ML})\|. \qquad (22)
\end{aligned}$$

On the other hand, one deduces from Assumptions 1 (i) that $\nabla_{10}I(\theta,\bar\theta)$ is locally Lipschitz in the variables θ and $\bar\theta$. Then, since, $\{\theta^k\}$ is bounded, there exists a bounded set $\mathcal{B}$ containing $\{\theta^k\}$ and a finite constant L such that for all θ, θ', $\bar\theta$ and $\bar\theta'$ in $\mathcal{B}$,

$$\|\nabla_{10}I(\theta,\bar\theta) - \nabla_{10}I(\theta',\bar\theta')\| \le L\big(\|\theta-\theta'\|^2 + \|\bar\theta-\bar\theta'\|^2\big)^{\frac{1}{2}}.$$

Using the triangle inequality and this last result, (22) asserts that for any $\theta \in \mathcal{B}$

$$\begin{aligned}
\beta_k L\|\theta^{k+1}-\theta_{ML}\| + \|R(\theta^{k+1}-\theta_{ML})\| \ge \|\big(\nabla^2 l_y(\theta_{ML}) \\
- \beta_k\nabla^2_{10}I(\theta_{ML},\theta_{ML})\big)(\theta^{k+1}-\theta_{ML})\|. \qquad (23)
\end{aligned}$$

Now, consider again the bounded set $\mathcal{B}$ containing $\{\theta^k\}$. Let λ_{l_y} and λ_I denote the minima

$$\lambda_{l_y} = \min_{\theta\in\mathcal{B}}\left\{-\lambda_{\nabla^2 l_y(\theta)}\right\}$$

$$\lambda_I = \min_{\theta,\bar\theta\in\mathcal{B}}\left\{\lambda_{\nabla^2_{10}I(\theta,\bar\theta)}\right\}.$$

Since for any symmetric matrix H, $x^T Hx/\|x\|^2$ is lower bounded by the minimum eigenvalue of H, we have immediately that

$$\begin{aligned}
&\|\big(-\nabla^2 l_y(\theta_{ML}) + \beta_k\nabla^2_{10}l_I(\theta_{ML},\theta_{ML})\big)(\theta^{k+1}-\theta_{ML})\|^2 \\
&\ge \big(\lambda_{l_y} + \beta_k\lambda_I\big)^2\|\theta^{k+1}-\theta_{ML}\|^2. \qquad (24)
\end{aligned}$$

By Assumptions 1.iii and 1.iv, $\lambda_{l_y} + \beta_k \lambda_I > 0$ and, after substitution of (24) into (23), we obtain

$$\beta_k L \|\theta^k - \theta_{ML}\| + \|R(\theta^{k+1} - \theta_{ML})\| \geq \\ \left(\lambda_{l_y} + \beta_k \lambda_I\right) \|\theta^{k+1} - \theta_{ML}\|, \tag{25}$$

for all $\theta \in \mathcal{B}$. Therefore, collecting terms in (25)

$$\beta_k L \geq \left(\lambda_{l_y} + \beta_k \lambda_I - \frac{\|R(\theta^{k+1} - \theta_{ML})\|}{\|\theta^{k+1} - \theta_{ML}\|}\right) \frac{\|\theta^{k+1} - \theta_{ML}\|}{\|\theta^k - \theta_{ML}\|}. \tag{26}$$

Now, recall that $\{\theta^k\}$ is convergent. Thus, $\lim_{k\to\infty} \|\theta^k - \theta_{ML}\| = 0$ and subsequently, $\lim_{k\to\infty} \frac{\|R(\theta^{k+1} - \theta_{ML})\|}{\|\theta^{k+1} - \theta_{ML}\|} = 0$ due to the definition of the remainder R. Since convergence of $\{\theta^k\}$ also implies that $\lim_{k\to\infty} \|\theta^k - \theta^{k-1}\| = 0$, invoking the assumptions of Lemma 3 yields: $\lim_{k\to\infty} \beta_k = \lim_{k\to\infty} \phi(\|\theta^k - \theta^{k-1}\|) = 0$. Therefore, as $\lambda_{l_y} > 0$, equation (26) gives

$$\lim_{k\to\infty} \frac{\|\theta^{k+1} - \theta_{ML}\|}{\|\theta^k - \theta_{ML}\|} = 0,$$

and the proof of superlinear convergence is completed.

Remark 2 *The convergence Theorems 1 and 2 make use of concavity of $l_y(\theta)$ and convexity of $I(\theta, \bar{\theta})$ via Assumptions 1.iii and 1.iv. However, for smooth non-convex functions an analogous local superlinear convergence result can be established under stronger assumptions similar to those in [**?**].*

5 Second order Approximations and Trust Region techniques

The main drawback of the Kullback proximal point algorithm (14) is that for $\beta_k \neq 1$ recursion (14) may be difficult to implement. In this section, we discuss an easily implementable version of the KPP algorithm using second order approximations which preserve monotonicity. The second order scheme is related to the well-known Trust Region technique for optimization introducedby Moré [**?**].

5.1 Approximate models

In order to obtain computable iterations, the following second order approximations of $l_y(\theta)$ and $I(\theta, \theta^k)$ can be introduced

$$\hat{l}_y(\theta) = l_y(\theta^k) + \nabla l_y(\theta^k)^\top(\theta - \theta^k) + \frac{1}{2}(\theta - \theta^k)^\top \nabla^2 l_y(\theta)(\theta - \theta^k).$$

and

$$\hat{I}(\theta, \theta^k) = \frac{1}{2}(\theta - \theta^k)^\top \nabla_{10}^2 I(\theta^k, \theta^k)(\theta - \theta^k).$$

In the following, we adopt the notations

$$g_k = \nabla l_y(\theta^k)$$
$$H_k = \nabla^2 l_y(\theta^k)$$

and

$$I_k = \nabla_{10}^2 I(\theta^k, \theta^k).$$

The approximate KPP algorithm is defined as

$$\theta^{k+1} = \text{argmax}_{\theta \in \mathbb{R}^p} \{ l_y(\theta^k) + g_k(\theta - \theta^k) + \frac{1}{2}(\theta - \theta^k)^\top H_k(\theta - \theta^k) - \frac{\beta_k}{2}(\theta - \theta^k)^\top I_k(\theta - \theta^k) \} \tag{27}$$

At this point it is important to make several comments. Notice first that for $\beta_k = 0$, $k = 1, 2, \ldots$, the approximate step (27) is equivalent to a Newton step. It is well known that Newton's method, also known as Fisher scoring in statistics, has superlinear asymptotic convergence rate but may diverge if not properly initialized. Therefore, at least for small values of the relaxation parameter β_k, the approximate PPA algorithm may fail to converge for the same reasons as for Newton's method. On the other hand, for $\beta_k > 0$ the term $-\frac{\beta_k}{2}(\theta - \theta^k)^\top I_k(\theta - \theta^k)$ penalizes the distance of the next iterate θ^{k+1} to the current iterate θ^k. Hence, we can interpret this term as a regularization which stabilizes the possibly divergent Newton algorithm without sacrificing its superlinear asymptotic convergence rate. By appropriate choice of $\{\beta_k\}$ the iterate θ^{k+1} can be forced to remain in a region around θ^k over which the quadratic model $\hat{l}_y(\theta)$ is accurate. This idea is very popular in optimization and is better known as the "Trust Region technique" [**?**][**?**].

5.2 Relationship to Trust Region Techniques

The Trust Region strategy proceeds as follows. The model $\hat{l}_y(\theta)$ is maximized in a ball $B(\theta^k, \delta) = \{\|\theta - \theta^k\|_{I_k} \leq \delta\}$ centered at θ^k where δ is a proximity control parameter which may depend on k, and where $\|a\|_{I_k} = a^\mathsf{T} I_k a$ is a norm; well defined due to positive definiteness of I_k (Assumption 1.iv). Given an iterate θ^k consider a candidate θ^δ for θ^{k+1} defined as the solution to the constrained optimization problem

$$\theta^\delta = \mathrm{argmax}_{\theta \in \mathbb{R}^p} \hat{l}_y(\theta)$$

subject to

$$\|\theta - \theta^k\|_{I_k} \leq \delta. \tag{28}$$

By duality theory of constrained optimization [?], and the fact that $\hat{l}_y(\theta)$ is strictly concave, this problem is equivalent to the unconstrained optimization

$$\theta^\delta(\beta) = \mathrm{argmin}_{\theta \in \mathbb{R}^p} L(\theta, \beta). \tag{29}$$

where

$$L(\theta, \beta) = -\hat{l}_y(\theta) + \frac{\beta}{2}\left(\|\theta - \theta^k\|_k^2 - \delta^2\right).$$

and β is a Lagrange multiplier selected to meet the constraint (28) with equality: $\|\theta^\delta(\beta) - \theta\|_{I_k} = \delta$.

We conclude that the Trust Region candidate θ^δ is identical to the approximate KPP iterate (27) with regularization parameter β chosen according to constraint (28). This relation also provides a rational rule for computing the relaxation parameter β.

5.3 Implementation

The parameter δ is said to be safe if θ^δ produces an acceptable increase in the original objective l_y. An iteration of the Trust Region method consists of two principal steps

Rule 1. Determine whether δ is safe or not. If δ is safe, set $\delta_k = \delta$ and take an approximate Kullback proximal step $\theta^{k+1} = \theta^\delta$. Otherwise, take a *null step* $\theta^{k+1} = \theta^k$.

Rule 2. Update δ depending on the result of *Rule 1.*

Rule 1 can be implemented by comparing the increase in the original log-likelihood l_y to a fraction m of the expected increase predicted by the approximate model $\hat{l}_y(\theta)$. Specifically, the Trust Region parameter δ is accepted if

$$l_y(\theta^\delta) - l_y(\theta^k) \geq m\big(\hat{l}_y(\theta^\delta) - \hat{l}_y(\theta^k)\big). \tag{30}$$

Rule 2 can be implemented as follows. If δ was accepted by Rule 1, δ is increased at the next iteration in order to extend the region of validity of the model $\hat{l}_y(\theta)$. If δ was rejected, the region must be tightened and δ is decreased at the next iteration.

The Trust Region strategy implemented here is essentially the same as that proposed by Moré [**?**].

Algorithm 1 *Step 0. (Initialization) Set* $\theta^0 \in \mathbb{R}^p$*,* $\delta_0 > 0$ *and the "curve search" parameters* m*,* m' *with* $0 < m < m' < 1$*.*

Step 1. Solve

$$\theta^{\delta_k} = \text{argmax}_{\theta\in\mathbb{R}^p}\hat{l}_y(\theta)$$

subject to

$$\|\theta - \theta^k\| \leq \delta_k.$$

Step 2. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \geq m\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ *then set* $\theta^{k+1} = \theta^{\delta_k}$*. Otherwise, set* $\theta^{k+1} = \theta^k$*.*

Step 3. Set $k = k+1$*. Update the model* $\hat{l}^k_y(\theta)$*. Update* δ_k *using Procedure 1.*

Step 4. Go to Step 1.

The procedure for updating δ_k is given below.

Procedure 1 *Step 0. (Initialization) Set* γ_1 *and* γ_2 *such that* $\gamma_1 < 1 < \gamma_2$*.*

Step 1. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \leq m\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ *then take* $\delta_{k+1} \in (0, \gamma_1\delta_k)$*.*

Step 2. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \leq m'\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ *then take* $\delta_{k+1} \in (\gamma_1\delta_k, \delta_k)$*.*

Step 3. If $l_y(\theta^{\delta_k}) - l_y(\theta^k) \geq m'\big(\hat{l}_y(\theta^{\delta_k}) - \hat{l}_y(\theta^k)\big)$ *then take* $\delta_{k+1} \in (\delta_k, \gamma_2\delta_k)$*.*

This algorithm satisfies the following convergence theorem

Theorem 3 *Let $g(y;\theta)$ and $k(x;\theta)$ be such that Assumptions 1 are satisfied. Then, $\{\theta^k\}$ generated by Algorithm 1 converges to the maximizer θ_{ML} of the log-likelihood $l_y(\theta)$ and satisfies the monotone likelihood property $l_y(\theta^{k+1}) \geq l_y(\theta^k)$. If in addition, the sequence of Lagrange multipliers $\{\beta_k\}$ tends towards zero, $\{\theta^k\}$ converges superlinearly.*

Remark 3 *The proof of Theorem 3 is omitted since it is standard in the analysis of Trust Region methods. Superlinear convergence as $\{\beta_k\}$ vanishes comes from the Dennis and More criterion [**?**, Theorem 3.11].*

Remark 4 *The Trust Region framework can also be applied to nonconvex contexts. In the case where I_k remains positive definite, global convergence to a local maximizer of $l_y(\theta)$ can be obtained under Assumptions 1.i and 1.ii only following the proof technique of [**?**].*

6 Application to Poisson data

In this section, we illustrate the application of Algoritm 1 for a maximum likelihood estimation problem in a Poisson inverse problem arising in emission computed tomography (ECT) such as positron emission tommography (PET).

6.1 The Poisson Inverse Problem

The objective is to estimate the intensity vector $\theta = [\theta_1, \ldots, \theta_p]^T$ governing the number of gamma-ray emissions $N = [N_1, \ldots, N_p]^T$ over an imaging volume of p pixels. The estimate of θ must be based on a vector of m observed projections of N denoted $Y = [Y_1, \ldots, Y_m]^T$. The components of N_i of N are independent Poisson distributed with rate parameters θ_i, and the components Y_j of Y are independent Poisson distributed with rate parameters $\sum_{i=1}^p P_{ji}\theta_i$, where P_{ji} is the transition probability corresponding to emissions from pixel i being detected at detector module j. The standard choice of complete data X, introduced by Shepp and Vardi [**?**], is the set $\{N_{ji}\}_{1\leq j\leq m,\ 1\leq i\leq p}$, where N_{ji} denotes the number of emissions in pixel i which are detected at detector j. The corresponding many-to-one mapping $h(X) = Y$ in the EM algorithm is

$$Y_j = \sum_{i=1}^{p} N_{ji}, \quad 1 \leq j \leq m. \tag{31}$$

It is also well known that the likelihood function is given by

$$\log g(y;\theta) = \sum_{j=1}^{m} \left(\sum_{i=1}^{p} P_{ji}\theta_i \right) - y_j \log \left(\sum_{i=1}^{p} P_{ji}\theta_i \right) + \log y_j! \tag{32}$$

and that the expectation step of the EM algorithm is (see [**?**])

$$Q(\theta,\bar{\theta}) = \mathsf{E}[\log f(x;\theta) \mid y;\bar{\theta}] = \tag{33}$$

$$\sum_{j=1}^{m}\sum_{i=1}^{p} \left(\frac{y_j P_{ji}\bar{\theta}_i}{\sum_{i=1}^{p} P_{ji}\bar{\theta}_i} \log(P_{ji}\theta_i) - P_{ji}\theta_i \right).$$

6.2 Approximate KPP using Trust Regions

The hessian $\nabla^2_{10} I(\theta,\bar{\theta})$ is obtained using the following straigtforward identity

$$\nabla^2_{10} I(\theta,\bar{\theta}) = \nabla^2_{10} Q(\theta,\bar{\theta}) - \nabla^2 l_y(\theta). \tag{34}$$

$$\nabla l_y(\theta^k) = \Big[g_i \Big]_{1\le i\le p} \tag{35}$$

with

$$g_i = \sum_{j=1}^{m} P_{ji}\theta_i^k - \frac{y_j P_{ji}}{\sum_{l=1}^{p} P_{jl}\theta_l^k} \tag{36}$$

and

$$\nabla^2 l_y(\theta^k) = [H_k^{ii'}]_{1\le i,i'\le p} \tag{37}$$

with

$$\begin{cases} H_k^{ii} = \sum_{j=1}^{m} \left(P_{ji} + \frac{y_j P_{ji}^2}{(\sum_{l=1}^{p} P_{jl}\theta_l^k)^2} \right), \\ H_k^{ii'} = \sum_{j=1}^{m} y_j \frac{P_{ji}P_{ji'}}{(\sum_{l=1}^{p} P_{jl}\theta_l^k)^2} \qquad i \neq i'. \end{cases} \tag{38}$$

The hessian of the $I(\theta,\bar{\theta})$ in (33) is

$$\nabla^2 \mathsf{E}[\log f(x;\theta) \mid y;\theta^k] = \Big[F_{ii'} \Big]_{1\le i,i'\le p} \tag{39}$$

with

$$\begin{cases} F_k^{ii} = -\frac{1}{\theta_i^{k\,2}} \sum_{j=1}^m \frac{y_j P_{ji} \theta_j^k}{\sum_{l=1}^p P_{jl} \theta_l^k}, \\ F_k^{ii'} = 0. \quad i \neq i' \end{cases} \tag{40}$$

Finally, using (34)

$$I_k = F_k - H_k. \tag{41}$$

6.3 Simulation results