Superlinearly Convergent Algorithms for ML Estimation: Proximal Point Methods, Kullback Regularization and Relaxed EM Steps

Stéphane Chrétien and Alfred O. Hero

In this paper, we study a sequential maximization method using the proximal point idea with Kullback-type regularization for computing maximum likelihood estimators. We prove that the EM algorithm belongs to the proposed class of algorithms and we introduce a relaxed version of the EM iterations. For an adequate choice of relaxation weights, superlinear convergence of the method is established. Then, implementable versions are proposed leading, in particular, to smoothened Newton and Quasi-Newton algorithms.

1 Introduction

Let $(\mathcal{X}, \mathcal{F}_X)$ and $(\mathcal{Y}, \mathcal{F}_Y)$ be two measurable spaces and let X and Y be random variables taking values on $\mathcal{X}$ and $\mathcal{Y}$ respectively. We postulate the family of densities $f(x;\theta)$ (resp. $g(y;\theta)$) defined on $\mathcal{X}$ (resp. $\mathcal{Y}$) parameterized by the deterministic vector $\theta \in R^p$. We also consider an additional family of probability measures $\kappa(x;\theta)$ defined on $\mathcal{F}_X$ which will have various interpretations in the sequel. Let μ be a measure on $\mathcal{F}_X$. Recall that two measures μ_1 and μ_2 are said to be absolutely continuous with respect to one-another if there is no set (event) $E \in \mathcal{F}_x$ such that $\mu_1(E) = 0$ and $\mu_2(E) \neq 0$ or $\mu_1(E) \neq 0$ and $\mu_2(E) = 0$, and we note $\mu_1 \equiv \mu_2$. In the sequel, we assume that $\kappa(\theta) \equiv \kappa(\theta')$ for all θ and θ' in R^p. This property is often called regularity of the familly $\kappa(\theta)$. Assume in addition that $\kappa(\theta) \equiv \mu$ for all θ in R^p. Then, for any $\theta \in R^p$, $k(x;\theta)$ denotes the Radon-Nikodym derivative of $\kappa(\theta)$ with respect to μ. In this paper, the densities $f(x;\theta)$, $g(y;\theta)$ and $k(x;\theta)$ are assumed twice differentiable in the variable θ. For any integrable function $F(x)$, $\mathsf{E}_\theta[F(x)]$ denotes the expectation $\int_{\mathcal{X}} F(x)k(x;\theta)d\mu(x)$. For any twice differentiable function $F(\theta)$, $\nabla F(\theta)$ denotes the gradient of F and $\nabla^2 F(\theta)$ the hessian of F, both taken at θ. Finally, for any function $F(\theta', \theta)$ of two variable vectors, twice differentiable in the second variable for a fixed θ', $\nabla F(\theta', \theta)$ and $\nabla^2 F(\theta', \theta)$ denote respectively the gradient and the hessian of $F(\theta', .)$ computed at θ.

The problem of maximum likelihood estimation consists in finding a solution to

$$\max_{\theta \in R^p} \log g(y;\theta), \tag{1}$$

where y is the vector of observed data. One of the most widely utilized methods for solving the ML estimation problem is the Expectation Maximization (EM) algorithm of Dempster*et* *al.*, which may be described as follows. First, $f(x;\theta)$ denotes the Radon-Nikodym derivative of a given probability measure with respect to μ. Then, it is assumed that the data y (denoted incomplete data) are obtained from the data x (denoted complete data) through a many-to-one mapping $x \mapsto y = h(x)$ so that the corresponding families of densities satisfy

$$g(y;\theta) = \int_{h^{-1}(\{y\})} f(x;\theta) d\mu(x) \tag{2}$$

In the framework of EM algorithms, $k(x;\theta)$ denotes the density of x conditionally to the observation y, i.e.

$$k(x;\theta) = \begin{cases} \frac{f(x;\theta)}{g(y;\theta)} & \text{if } x \in h^{-1}(\{y\}), \\ 0 & \text{otherwise .} \end{cases}$$

Notice that, due to relation (2), the density $k(x;\theta)$ is always defined. Then, the EM algorithm generates a sequence starting from an initial point $\theta^0 \in R^p$ and defined by

$$\begin{aligned} &\text{Compute } \mathsf{E}[\log f(x;\theta)|y,\theta^k], && (\ \textbf{E step}\ ) \\ &\theta^{k+1} = \text{argmax}_{\theta \in R^p} \mathsf{E}[\log f(x;\theta)|y,\theta^k] && (\ \textbf{M step}\ ) \end{aligned} \tag{3}$$

Due to its simplicity of implementation and its safe monotonic convergence property, the EM algorithm enjoyed increasing interest during the past two decades. Multiple strategies have been proposed for improving the speed of convergence [8, 2] and the behaviour of the method seems to be well understood at present [11].

The objective of this paper is to address more general approaches for the ML estimation problem and study their convergence. We propose a method which finds an adequate formulation in terms of the proximal point method of Rockafellar [10] using a Kullback regularization of the log −likelihood function. The standard EM algorithm is shown to belong to this new methodology. Our approach opens up many perspectives for ML estimation. In particular, we study a relaxed version of the EM algorithm. The main consequence of our approach is contruction of superlinearly convergent methods with adequate choice of parameters. In particular, our relaxed version of the EM algorithm is superlinearly convergent. We then study approximations of the method with low numerical complexity. Finally, we show application of the proposed relaxed EM algorithms to the case of Poisson data as it occurs in PET image reconstruction.

2 Kullback regularization and proximal methods

In this section, the proximal point strategy is introduced. We then propose a proximal method using Kullback regularization for ML estimation.

2.1 The proximal point approach

The proximal point algorithm introduced by Rockafellar [10] is a method of sequential maximizations for solving the general optimization problem

$$\max_{\theta \in R^p} \Phi(\theta),$$

where $\Phi(\theta)$ is a concave function. Consider the gradient operator associated to Φ, i.e.

$$T: \quad \theta \mapsto T(\theta) = \nabla\Phi(\theta).$$

Finding a maximum of the concave function Φ is thus equivalent to finding a point for which $T(\theta) = 0$. In the case of a nonconcave function, a point satisfying $T(\theta) = 0$ is called a stationary point and may be a maximum, a minimum of a saddle point. To overcome this difficulty, local concavity about the solution of the ML estimation problem is often required in the study of EM-type algorithms. For a sake of clarity, assume that Φ is concave, bounded from above and that $\mathrm{dom}(\Phi) = R^p$, where $\mathrm{dom}(\Phi) = \{\theta \in R^p \quad | \quad -\Phi(\theta) < +\infty$. The proximal point algorithm is a method of successive approximation of the fixed point of the operator

$$P_\lambda = (\mathrm{Id} + \frac{1}{2\lambda}T)^{-1},$$

defined over R^p, thanks to the prescribed assumptions, where Id denotes identity and λ is a positive real number. Indeed, a closer look at this last formula shows that

$$\begin{aligned} \theta - P_\lambda(\theta) &= \frac{1}{2\lambda} T(P_\lambda(\theta)) \\ &= \frac{1}{2\lambda} \nabla\Phi(P_\lambda(\theta)), \end{aligned} \tag{4}$$

and therefore, any fixed point of P_λ satisfies $\nabla\Phi(\theta) = 0$ as desired. Moreover, P_λ can clearly be written

$$P_\lambda(\theta) = \mathrm{argmax}_{\theta' \in R^p}\{\Phi(\theta') - \lambda|\theta - \theta'|^2\}, \tag{5}$$

since the optimality condition on the right hand side is exactly (4). The wording "proximal" comes from the fact that if $\Phi(\theta)$ is the indicator function of some closed set S, then $P_\lambda(\theta)$ is the projection of θ onto S. With these definitions in hand, the proximal point algorithm is the straightforward application of the well know successive approximation method of fixed points [] defined by the recurrence

$$\theta^{k+1} = P_\lambda(\theta^k),$$

given the initial point θ^1. Using the formulation of (5) and allowing the relaxation parameter to vary with iterations, we obtain

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in R^p} \Phi(\theta) - \lambda_k|\theta^k - \theta|^2. \tag{6}$$

This last equation expresses the maximization problem in terms of successive maximization subproblems regularized by a quadratic penalty. This regularization if often known as Moreau-Yosida regularization [9]. Finally, since iteration (6) may not have a computable solution, the proximal point scheme is generally approximated, for instance by using polyhedral models of function Φ, and one obtains a implementable version of the proximal point algorithm, that we will simply write

$$\theta^{k+1} \approx \text{argmax}_{\theta \in R^p} \Phi(\theta) - \lambda_k |\theta^k - \theta|^2. \tag{7}$$

For the concern of our study, it is fundamental to underline that regularization by a quadratic functional is far from beeing mandatory and should be advantageously replaced in (6) by other meaningfull penalties. In the area of ML estimation, statistical distances offer a much more appropriate class of regularizations, as will be proved in the sequel. Our main focus in the present work is the use of Kullback regularization in the proximal point framework.

We would like to end this presentation of the proximal point approach by emphasizing the fact that proximal approaches form actually the basis of a large number of recent optimization techniques [7].

2.2 Kullback regularization of proximal methods for ML estimation

The main results of this paper concern the study of the proximal point method using Kullback information measures as regularization. The procedure is directly obtained from the standard proximal scheme (6) and is given here in the case of ML estimation,

$$\theta^{k+1} = \text{argmax}_{\theta \in R^p} \log g(y;\theta) - \lambda_k I(\theta^k, \theta), \tag{8}$$

where $(\lambda_k)_{k \in N}$ is a sequence of positive weights, and $I(\theta', \theta)$ is the Kullback distance defined by

$$I(\theta', \theta) = \mathsf{E}_{\theta'}[\log \frac{k(x;\theta')}{k(x;\theta)}].$$

Notice that, since we assumed $\kappa(\theta') \equiv \kappa(\theta)$ for all θ' and θ in R^p, this last quantity is well defined altough it may take infinite values [6]. In addition, recall that $I(\theta', \theta) \geq 0$, and that clearly $I(\theta, \theta) = 0$ for all θ' and θ in R^p. Thus, one can check that $\nabla I(\theta, \theta) = 0$ for all $theta \in R^p$, since $I(\theta', \theta)$ is positive and attains its minimum at zero.

In this framework, $k(x;\theta)$ is any density on $\mathcal{X}$, but can denote in particular the density of X conditionally to the observed data y as in the case of the EM algorithm. In order to validate our methodology, one must prove that, as in the standard proximal point algorithm, fixed points of procedure (8) are stationary points of the objective function (e.g. the likelihood function). Thus, let us consider the first order optimality condition at a fixed point θ^* of the procedure. For a sake of simplicity, fix $\lambda_k = \lambda$ in (8). Then, the parameter θ^* is a fixed point of iteration (8) if

$$0 = \nabla \log g(y;\theta^*) - \lambda \nabla I(\theta^*, \theta^*), \tag{9}$$

where $\nabla I(\theta^*, \theta^*)$ denotes the gradient of $I(\theta^*, .)$ at θ^*. Recall that $\nabla I(\theta, \theta) = 0$ for any parameter θ. Thus, equation (9) gives

$$0 = \nabla \log g(y;\theta^*), \tag{10}$$

and θ^* is a stationary point of $\log g(y;\theta)$, as desired. Finally existence of a maximizer in iteration (8) is necessary to computation. This is obtained if the following assumptions are satisfied

Assumptions 1.

(*i*) $\mathrm{dom}(-g(y;\theta)) = R^p$ *and* $\lim_{|\theta|\to\infty} \log g(y;\theta) = -\infty$,
(*ii*) $\mathrm{dom}(I(\theta',\theta)) = R^p$ *and* $\lim_{|\theta|\to\infty} I(\theta',\theta) = \infty$ *for all* $\theta' \in R^p$.

Then, the following definition summarizes the proximal approach for ML estimation.

Definition 1. Kullback proximal point algorithm for ML
Let $(\lambda_k)_{k\in N}$ *be a sequence of positive relaxation weights. Let assumptions 1 be satisfied. Then, the proximal point algorithm using Kullback regularization (PPAK) for ML estimation is defined, given a starting point* θ^1*, by the recurrence*

$$\theta^{k+1} = \mathrm{argmax}_{\theta\in R^p} \log g(y;\theta) - \lambda_k I(\theta^k,\theta). \tag{11}$$

Finally, we would like to point out that several alternatives to the quadratic (Moreau-Yosida) functional have already been proposed in other works. Surprisingly, all the former studies on the subject are strongly related to the concept of entropy, even in nonstatistical contexts. We recommend the study in [4] for further information. However, the use of Kullback information measures as proposed in this paper appears for the first time in proximal methods and is original in its statistical interpretations for ML estimation. We now establish the relationships between the well known EM algorithm and our method.

3 Relaxation of the EM algorithm

In this section, we redefine the EM algorithm as a particular instance of the proximal point algorithm with Kullack relaxation (PPAK). Moreover, using this framework, we obtain a relaxed version of the EM procedure, the relaxation parameter having a great potential for the control of the speed of convergence as studied in section 4 below.

Relationships between the EM algorithm and the proximal point approach have never been clearly established in the past, although some links wich Bregman regularization were pointed out in the special case of Poisson variables in [5] (see [4] for the links between Bregman functions and the proximal point approach). The following proposition clarifies this important fact.

Proposition 1. *Given the observation data y, the EM algorithm is a proximal point algorithm using Kullback regularization (PPAK) applied to the likelihood function* $\log g(y;\theta)$*, in the particular case of unit relaxation and* $I(\theta',\theta) = \mathsf{E}_{\theta'}[\log \frac{k(y;\theta')}{k(y;\theta)}]$*, with*

$$k(x;\theta) = \begin{cases} \frac{f(x;\theta)}{g(y;\theta)} & \textit{if } x \in h^{-1}(\{y\}), \\ 0 & \textit{otherwise} . \end{cases}$$

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

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

which gives

$$\theta^{k+1} = \text{argmax}_{\theta \in R^p} \log g(y;\theta) - \mathsf{E}_{\theta^k}[\log \frac{k(x;\theta^k)}{k(x;\theta)}] \qquad (\textbf{ M step }). \tag{12}$$

We thus recover the formulation of a proximal point algorithm with relaxation weight $\lambda_k = 1$. $\square$

Using this proposition and the definition of the proximal point method with Kullback regularization, it is immediate to generalize the EM algorithm to the case of relaxed iterations.

Definition 2. Relaxed EM algorithm
Let $(\lambda_k)_{k \in N}$ be a sequence of positive relaxation weights. Then, given the observation data y, the Relaxed EM algorithm is defined by the following recurrence,

$$\begin{aligned} &\text{Compute } \mathsf{E}_{\theta'}[\log \tfrac{k(x;\theta')}{k(x;\theta)}], && (\textbf{ E step }) \\ &\theta^{k+1} = \text{argmax}_{\theta \in R^p} \log g(y;\theta) - \lambda_k \mathsf{E}_{\theta^k}[\log \tfrac{k(x;\theta^k)}{k(x;\theta)}] && (\textbf{ Relaxed M step }) \end{aligned} \tag{13}$$

with

$$k(x;\theta) = \begin{cases} \frac{f(x;\theta)}{g(y;\theta)} \text{ if } x \in h^{-1}(\{y\}), \\ 0 \text{ otherwise }. \end{cases}$$

Alternatively the Relaxed EM algorithm can be written in the equivalent form

$$\begin{aligned} &\text{Compute } \mathsf{E}_{\theta^k}[\log f(x;\theta)], && (\textbf{ E step }) \\ &\theta^{k+1} = \text{argmax}_{\theta \in R^p}(1-\lambda_k) \log g(y;\theta) + \lambda_k \mathsf{E}_{\theta^k}[\log f(x;\theta)] && (\textbf{ Relaxed M step }) \end{aligned} \tag{14}$$

Notice that this form enjoys the nice property of simply requiring the computation of the standard E step quantity $\mathsf{E}_{\theta^k}[\log f(x;\theta)]$. We now prove superlinear convergence of the proximal point algorithm using Kullback regularization.

4 Convergence

In this section, we provide strong convergence results of the proximal point algorithm using Kullback regularization. In particular, we show that for a particular choice of the sequence of relaxation weights λ_k, the method is superlinearly convergent towards a maximizer. Notice that such a strong convergence result has, to our knowledge, no precedent in the EM litterature.

In the sequel, we will often use the concept of elliptic functions. Recall that a function $\Phi(\theta)$ is said to be locally elliptic if Φ is twice continuously differentiable and its hessian is positive definite over any bounded set $\mathcal{B}$. In particular, ellipticity implies the existence of constant numbers $0 < l(\mathcal{B}) \leq L(\mathcal{B})$ such that

$$l(\mathcal{B})|d|^2 \leq \langle \nabla^2 \Phi(\theta) d, d \rangle \quad \text{and} \quad |\nabla^2 \Phi(\theta) d| \leq |d|, \tag{15}$$

for all $d \in R^p$. In other words, $l(\mathcal{B})$ and $L(\mathcal{B})$ are respectively lower and upper bounds of the eigenvalues of the hessian matrix $\nabla^2\Phi(\theta)$ over $\mathcal{B}$.

The necessary assumptions for the convergence analysis are summarized below

Assumptions 2. *(i) The likelihood function satisfies* $\mathrm{dom}(\log g(y;\theta)) = R^p$ *and* $\lim_{|\theta|\to\infty} \log g(y;\theta) = -\infty$
(ii) $I(\theta',\theta)$ *is twice differentiable in both variables and is elliptic in the second variable for any* θ'
(iii) The sequence $(\lambda_k)_{k\in N}$ *is defined by* $\lambda_k = \lambda(|\theta^k - \theta^{k+1}|)$*, where* $\lambda(.)$ *is such that:*
a. $\lambda(.)$ *is continuous over* R*,*
b. $\lambda(t) \geq 0$ *for all* $t \in R$*,*
c. $\lambda(t) = 0 \Leftrightarrow t = 0$*.*

It is clear from assumption2 (ii) that for any $\theta' \in R^p$, $\lim_{|\theta|\to\infty} I(\theta',\theta) = \infty$ and thus, recurrence (11) is well defined. Furthermore, ellipticity of $I(\theta',\theta)$ over R^p in the second variable implies that if $\theta' \neq \theta$, then $I(\theta',\theta) > 0$. Notice also that if $\lim k \to \infty|\theta^{k+1} - \theta^k| = 0$, then $\lim_{k\to\infty} \lambda_k = 0$. We begin by proving the following usefull result.

Lemma 1. *Let assumptions 2 be satisfied. Then, given two bounded sequences* $(\theta_1^k)_{k\in N}$ *and* $(\theta_2^k)_{k\in N}$*,* $\lim_{k\to\infty} I(\theta_1^k,\theta_2^k) = 0$ *implies* $\lim_{k\to\infty} |\theta_1^k - \theta_2^k| = 0$*.*

Proof. It is well known that ellipticity of a function over a bounded set $\mathcal{B}$ implies strong convexity over this set. Moreover, for any bounded set $\mathcal{B}$ and $\theta' \in \mathcal{B}$, $I(\theta',\theta)$ is strongly convex on $\mathcal{B}$ in the second variable, i.e.

$$I(\theta',\theta_2) \geq I(\theta',\theta_1) + \langle \nabla I(\bar{\theta},\theta_1), \theta_2 - \theta_1\rangle + \frac{1}{2} l(\mathcal{B})|\theta_2 - \theta_1|^2. \tag{16}$$

Apply this formula with the iterates θ_1^k and θ_2^k to obtain

$$I(\theta_1^k,\theta_2^k) \geq I(\theta_1^k,\theta_1^k) + \langle \nabla I(\theta_1^k,\theta_1^k), \theta_2^k - \theta_1^k\rangle + \frac{1}{2} l(\mathcal{B})|\theta_2^k - \theta_1^k|^2. \tag{17}$$

Recall that $I(\theta,\theta) = 0$ and $\nabla I(\theta,\theta) = 0$ for any θ in $\mathcal{B}$ to deduce from (17) that

$$I(\theta_1^k,\theta_2^k) \geq \frac{1}{2} l(\mathcal{B})|\theta_2^k - \theta_1^k|^2.$$

The desired result is obtained by passing at the limit in this last equation. □

The following lemma is also important

Lemma 2. *Let* $(\theta^k)_{k\in N}$ *be a sequence defined by recurrence (11), with assumptions 2. Then the sequence* $(\theta^k)_{k\in N}$ *is bounded.*

Proof. Notice that, due to (11), we have

$$\log g(y;\theta^{k+1}) - \log g(y;\theta^k) \geq \lambda_k I(\theta^k,\theta^{k+1}) \ \ (\geq 0). \tag{18}$$

Thus, the sequence $(\log g(y;\theta^k))_{k\in N}$ is increasing. Now, recall that, due to assumption 2 (i), $\lim_{|\theta|\to\infty} \log g(y;\theta) = -\infty$. Then, the monotone increasing property of $(\log g(y;\theta^k))_{k\in N}$ implies that $(\theta^k)_{k\in N}$ is bounded using a straightforward argument by contradiction. □

We now prove the following important property of the sequence defined by the proximal point algorithm with Kullback regularization.

Lemma 3. *Let $(\theta^k)_{k\in N}$ be a sequence defined by recurrence (11), with assumptions 2. Then* $\lim_{k\to\infty} |\theta^{k+1} - \theta^k| = 0$.

Proof. Recall that equation (11) implies

$$\log g(y;\theta^{k+1}) - \log g(y;\theta^k) \geq \lambda_k I(\theta^k, \theta^{k+1}) \ (\geq 0). \tag{19}$$

as in the proof of lemma 2. Since $(\theta^k)_{k\in N}$ is bounded, there exists a closed bounded set $\mathcal{B}$ which contains the sequence $(\theta^k)_{k\in N}$. Notice that the likelihood function $\log g(y;\theta^k)$ is clearly bounded from above, thus, $(\log g(y;\theta^k))_{k\in N}$ is convergent and, we deduce from (19) that $\lim_{k\to\infty} \lambda(|\theta^k - \theta^{k+1}|)I(\theta^k,\theta^{k+1}) = 0$. Therefore, either $\lambda(|\theta^k - \theta^{k+1}|)$ or $I(\theta^k,\theta^{k+1})$ tends toward zero. In the first case, assumptions 2 (iii) imply that $\lim_{k\to\infty} |\theta^k - \theta^{k+1}| = 0$. In the second case, we easily conclude using lemma 1. □

We now establish stationarity of the accumulation points of the sequence of iterates.

Lemma 4. *Any accumulation point θ^* of the sequence $(\theta^k)_{k\in N}$ generated by recurrence (11), with assumptions 2, is a stationary point of* $\log g(y;\theta)$.

Proof. Consider a subsequence $(\theta^{\sigma(k)})_{k\in N}$ converging towards θ^*. Due to lemma 3, the sequence $(\theta^{\sigma(k)+1})_{k\in N}$ also converges towards θ^*. Now, recall that iteration (11) implies

$$\nabla(\log g(y;\theta^{\sigma(k)+1})) + \lambda_k \nabla I(\theta^{\sigma(k)}, \theta^{\sigma(k)+1}) = 0. \tag{20}$$

Twice continuous differentiability of $I(\theta',\theta)$ in both variables (assumption 2 (ii)) implies that $\nabla I(\theta',\theta)$ is a continuous function of (θ',θ), which implies that $\lim_{k\to\infty} \nabla I(\theta^{\sigma(k)},\theta^{\sigma(k)+1}) = \nabla I(\theta^*,\theta^*)$. Now, recall that $\nabla I(\theta^*,\theta^*) = 0$, so that equation (20) gives $\lim_{k\to\infty} \nabla(\log g(y;\theta^{\sigma(k)+1})) = 0$. Twice continuous differentiability of $g(y;\theta)$ and assumption 2 (i) imply that $\nabla(\log g(y;\theta^{\sigma(k)+1}))$ is continuous over each bounded set. Since $(\theta^{\sigma(k)+1})_{k\in N}$ is bounded and converges to θ^*, we finally obtain $\nabla(\log g(y;\theta^*)) = 0$ as desired. □

We are now ready to establish our convergence theorem. For that purpose, we define the following level set

$$\mathcal{L}_{\theta^1} = \{\theta \in R^p \quad | \quad \log g(y;\theta) \geq \log g(y;\theta^1)\}$$

Theorem 1. *Consider the sequence $(\theta^k)_{k\in N}$ generated by recurrence (11). with assumptions 2. In addition, assume that $-\log g(y;\theta)$ is locally elliptic over $\mathcal{L}_{\theta^1}$. Then $(\theta^k)_{k\in N}$ converges superlinearly to the unique maximizer of* $\log g(y;\theta)$.

Proof. First, we prove convergence of the sequence $(\theta^k)_{k\in N}$ to the optimum. We assume in the following that θ^1 is not a stationary point of $\log g(y;\theta)$. Recall that $(\theta^k)_{k\in N}$ is bounded due to lemma 2. Let θ^* be an accumulation point of the sequence $(\theta^k)_{k\in N}$. Due to lemma 4, θ^* is a stationary point of $\log g(y;\theta)$. Since θ^1 is not stationary, $\theta^* \neq \theta^1$ and thus, there exists an iteration number k such that $\theta^{k+1} \neq \theta^k$. Then, ellipticity of $I(\theta', \theta)$ in the second variable implies that $I(\theta^{k+1}, \theta^k) > 0$. Therefore, (19) clearly implies that $\log g(y;\theta^*) > \log g(y;\theta^1)$ and thus, θ^* lies in the interior of $\mathcal{L}_{\theta^1}$. Now recall that $-\log g(y;\theta)$ is elliptic over $\mathcal{L}_{\theta^1}$. Thus, $-\log g(y;\theta)$ is strictly convex over $\mathcal{L}_{\theta^1}$, which implies that θ^* is the unique stationary point of $\log g(y;\theta)$. Thus, the existence of an accumulation point θ' of $(\theta^k)_{k\in N}$ different from θ^* is impossible due lemma 4. Since the sequence of iterates $(\theta^k)_{k\in N}$ has only one accumulation point θ^*, it converges towards θ^*.

We now prove superlinear convergence of the sequence of iterates towards the optimum parameter θ^*. The gradient function $\nabla g(y;\theta) - \lambda_k \nabla I(\theta^*, \theta)$ has the following Taylor expansion at the point θ^*,

$$\begin{aligned}\nabla \log g(y;\theta) - \lambda_k \nabla I(\theta^*, \theta) =& \nabla \log g(y;\theta^*) - \lambda_k \nabla I(\theta^*, \theta^*) + \nabla^2 \log g(y;\theta^*)(\theta - \theta^*) \\ &- \lambda_k \nabla^2 I(\theta^*, \theta^*)(\theta - \theta^*) + R(\theta - \theta^*),\end{aligned} \tag{21}$$

with $\lim_{k\to\infty} \frac{|R(\theta-\theta^*)|}{|\theta-\theta^*|} = 0$. Since θ^* is a stationary point of $\log g(y;\theta)$, we have $\nabla \log g(y;\theta^*) = 0$ and recalling that $\nabla I(\theta^*, \theta^*) = 0$, we obtain, taking $\theta = \theta^{k+1}$,

$$\begin{aligned}\nabla \log g(y;\theta^{k+1}) - \lambda_k \nabla I(\theta^*, \theta^{k+1}) =& \nabla^2 \log g(y;\theta^*)(\theta - \theta^*) - \lambda_k \nabla^2 I(\theta^*, \theta^*)(\theta^{k+1} - \theta^*) \\ &+ R(\theta^{k+1} - \theta^*).\end{aligned} \tag{22}$$

Recall that, due to iteration (11),

$$\nabla g(y;\theta^{k+1}) - \lambda_k \nabla I(\theta^k, \theta^{k+1}) = 0. \tag{23}$$

Combining this last equation with (22) gives

$$\lambda_k\big(I(\theta^k, \theta^{k+1}) - I(\theta^*, \theta^{k+1})\big) = \nabla^2 \log g(y;\theta^*)(\theta - \theta^*) - \lambda_k \nabla^2 I(\theta^*, \theta^*)(\theta^{k+1} - \theta^*) + R(\theta^{k+1} - \theta^*), \tag{24}$$

from which we deduce that

$$\begin{aligned}&|\lambda_k\big(I(\theta^k, \theta^{k+1}) - I(\theta^*, \theta^{k+1})\big) - R(\theta^{k+1} - \theta^*)| \\ &\qquad = |\big(\nabla^2 \log g(y;\theta^*) - \lambda_k \nabla^2 I(\theta^*, \theta^*)\big)(\theta^{k+1} - \theta^*)|.\end{aligned} \tag{25}$$

On the other hand, since $I(.,.)$ is twice continuously differentiable over $R^p \times R^p$, it is clear that $I(., \theta^{k+1})$ is bounded over each bounded set $\mathcal{B}$ and is a locally Lipschitz function in the first variable, i.e. if θ and θ' belong to $\mathcal{B}$, there exists a constant $L_{\theta^{k+1}}(\mathcal{B})$ such that

$$|I(\theta, \theta^{k+1}) - I(\theta', \theta^{k+1})| \leq L_{\theta^{k+1}}(\mathcal{B})|\theta - \theta'|. \tag{26}$$

Now, since $\mathcal{B}$ is bounded and $I(.,.)$ is continuous and bounded over $\mathcal{B}$, the supremum over $\theta^{k+1} \in \mathcal{B}$ of $L_{\theta^{k+1}}(\mathcal{B})$ is finite and we have

$$|I(\theta, \theta^{k+1}) - I(\theta', \theta^{k+1})| \leq L_{\sup}(\mathcal{B})|\theta - \theta'|. \tag{27}$$

Now, since the sequence $(\theta^k)_{k\in N}$ belongs to $\mathcal{L}_{\theta^1}$ and $\mathcal{L}_{\theta^1}$ is bounded due to assumption 2, this last property implies that

$$|I(\theta^k,\theta^{k+1}) - I(\theta^*,\theta^{k+1})| \leq L_{\sup}(\mathcal{L}_{\theta^1})|\theta^k - \theta^*|. \tag{28}$$

Then, equation (25) gives

$$\lambda_k L_{sup}(\mathcal{L}_{\theta^1})|\theta^k - \theta^*| + |R(\theta^{k+1} - \theta^*)| \geq |\big(\nabla^2 \log g(y;\theta^*) - \lambda_k \nabla^2 I(\theta^*,\theta^*)\big)(\theta^{k+1} - \theta^*)|. \tag{29}$$

Furthermore, recall that $-\log g(y;\theta)$ is elliptic and thus, there exits a constant $l_g(\mathcal{L}_{\theta^1})$ such that

$$\langle d, \nabla^2 \log g(y;\theta) d\rangle \geq l_g(\mathcal{L}_{\theta^1})|d|^2, \tag{30}$$

for all $d \in R^p$ (see [1]). Therefore, combining this inequality with assumption 2, we obtain

$$\langle d, \Big(-\nabla^2 \log g(y;\theta) + \lambda_k \nabla^2 I(\theta^*,\theta^*)\Big) d\rangle \geq (l_g(\mathcal{L}_{\theta^1}) + \lambda_k l(\mathcal{L}_{\theta^1}))|d|^2, \tag{31}$$

and thus, applying Cauchy Schwartz inequality,

$$|\Big(\nabla^2 \log g(y;\theta) - \lambda_k \nabla^2 I(\theta^*,\theta^*)\Big) d| \geq (l_g(\mathcal{L}_{\theta^1}) + \lambda_k l(\mathcal{L}_{\theta^1}))|d| \tag{32}$$

for all $d \in R^p$. Using this last result, equation (29) becomes

$$\lambda_k L_{sup}(\mathcal{L}_{\theta^1})|\theta^k - \theta^*| + |R(\theta^{k+1} - \theta^*)| \geq \Big(l_g(\mathcal{L}_{\theta^1}) + \lambda_k l(\mathcal{L}_{\theta^1})\Big)|\theta^{k+1} - \theta^*|. \tag{33}$$

and we further obtain

$$\lambda_k L_{sup}(\mathcal{L}_{\theta^1})|\theta^k - \theta^*| \geq \Big(|l_g(\mathcal{L}_{\theta^1}) + \lambda_k l(\mathcal{L}_{\theta^1})| - \frac{|R(\theta^{k+1} - \theta^*)|}{|\theta^{k+1} - \theta^*|}\Big)|\theta^{k+1} - \theta^*|. \tag{34}$$

Now, recall that $\lim_{k\to\infty} \frac{|R(\theta^{k+1}-\theta^*)|}{|\theta^{k+1}-\theta^*|} = 0$. Hence, there exists a constant K such that $k \geq K$ implies

$$|l_g(\mathcal{L}_{\theta^1}) + \lambda_k l(\mathcal{L}_{\theta^1})| - \frac{|R(\theta^{k+1} - \theta^*)|}{|\theta^{k+1} - \theta^*|} \geq \frac{l_g(\mathcal{L}_{\theta^1})}{2}. \tag{35}$$

We then deduce from (34) that for k large enough,

$$\frac{|\theta^{k+1} - \theta^*|}{|\theta^k - \theta^*|} \leq \frac{2}{l_g(\mathcal{L}_{\theta^1})} \lambda_k L_{sup}(\mathcal{L}_{\theta^1}). \tag{36}$$

Finally, since $\lim_{k\to\infty} \lambda_k = 0$, we obtain

$$\lim_{k\to\infty} \frac{|\theta^{k+1} - \theta^*|}{|\theta^k - \theta^*|} = 0, \tag{37}$$

and the proof is completed. □

5 Implementation

In this section, we focus on implementable forms of the proximal point algorithm with Kullback regularization. Indeed, as observed in section 2.1, recurrence (11) may be difficult to compute in practice, although utilization of EM algorithms relies on the fact that exact computation can be achieved in the case $\lambda_k = 1$ for all k. In the following study, we present efficient algorithmic schemes which accurately approximate the proximal point iteration and preserve superlinear convergence.

5.1 Approximate iterations

Our approximation schemes are based on second order expansions of the likelihood and the Kullback regularization function. At iteration k, we introduce the following second order approximation of the likelihood,

$$\log g(y;\theta) \approx \log g(y;\theta^k) + \langle \nabla \log g(y;\theta^k), \theta - \theta^k \rangle + \langle \theta - \theta^k, G_k(\theta - \theta^k) \rangle,$$

and the regularization, recalling that $I(\theta^k,\theta^k) = 0$ and $\nabla I(\theta^k,\theta^k) = 0$

$$I(\theta^k,\theta) \approx \langle \theta - \theta^k, I_k\theta - \theta^k \rangle,$$

where the matrices G_k and I_k may correspond to various computational options. Thus, the following approximate version of the proximal point iteration is obtained

$$\theta^{k+1} = \mathrm{argmax}_{\theta \in R^p} \{ \ \langle \nabla \log g(y;\theta^k), \theta - \theta^k \rangle + \langle \theta - \theta^k, (G_k - \lambda_k I_k)(\theta - \theta^k) \rangle \ \}, \tag{38}$$

which is equivalent to the following recurrence

$$\theta^{k+1} = \theta^k - (G_k - \lambda_k I_k)^{-1} \nabla \log g(y;\theta^k). \tag{39}$$

This algorithmic pattern is quite general and allows various choices for the matrices G_k and I_k. In the following section, we investigate the details of two different implementations. Then, we prove superlinear convergence of one of them.

5.1.1 Second order approximation

In this approach, exact second order approximations are achieved. The matrices G_k and I_k take the values

$$\begin{aligned} G_k &= \nabla^2 \log g(y;\theta^k) \\ I_k &= \nabla^2 I(\theta^k,\theta^k). \end{aligned} \tag{40}$$

It is fundamental to notice that in this case, setting $\lambda_k = 0$ makes the approximate proximal algorithm equivalent to Newton's method. In the relaxed case, i.e. $\lambda_k \neq 0$, we obtain a corrected Newton step. The correction may vanish in the neighborhood of an optimal point and almost Newton steps may thus be taken. On the other hand, correction of Newton steps can be used to smoothen the trajectory of the iterates far from the solution since, in the proximal approach, I_k penalizes the distance between to successive iterates.

5.1.2 Quasi-second order approximation

A quasi-second order approach can be implemented in order to reduce numerical complexity. Let G_k and I_k be defined as in the preceding section, i.e. G_k is the hessian of the likelihood function and I_k is the hessian of the Kullback distance, both computed at iteration k. The inverse matrix $(G_k - \lambda_k I_k)^{-1}$ can be approximated by

$$(G_k - \lambda_k I_k)^{-1} \approx G_k^{-1} + \lambda_k G_k^{-1} I_k G_k^{-1}, \tag{41}$$

for small values of the relaxation parameter λ_k. Notice that this approach is consistent with theorem 1 where superlinear convergence is obtained with $\lim_{k\to\infty} \lambda_k = 0$. Then, the inverse of the hessian matrix G_k^{-1} of the likelihood function is iteratively approximated using BFGS update rule (see [1] for details)

$$G_k^{-1} \approx W_k, \tag{42}$$

with W_k defined by the following recurrence

$$W_{k+1} = W_k - \frac{s_{k+1} y_{k+1}^T M_k + W_k y_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}} + \left(1 + \frac{y_{k+1}^T W_k y_{k+1}}{y_{k+1}^T s_{k+1}}\right) \frac{s_{k+1} s_{k+1}^T}{y_{k+1}^T s_{k+1}},$$

taking

$$\begin{aligned} s_{k+1} &= \theta^{k+1} - \theta^k \\ y_{k+1} &= \nabla \log g(y; \theta^{k+1}) - \nabla \log g(y; \theta^k). \end{aligned}$$

This method clearly leads to a smoothened quasi-Newton algorithm.

5.2 Convergence

In this section, we prove convergence of the exact second order method of section 5.1.1. The quasi second order method necessitates more room than the scope of this paper permits and will be studied in a forthcoming paper. The following result shows that superlinear convergence to a maximizer is also satisfied by the approximated second order method.

Theorem 2. *Assume the existence of a maximizer θ^* in the interior of a bounded neighborhood $\mathcal{B}$ of a certain iterate θ^k and that $-\log g(y;\theta)$ is locally elliptic over $\mathcal{B}$. Assume also that $I(\theta', \theta)$ is elliptic over $\mathcal{B}$ for any θ' in $\mathcal{B}$. In addition, assume that $\lim_{k\to\infty} \lambda_k = 0$ Then, if θ^k is sufficiently close to θ^*, the sequence $(\theta^k)_{k\in N}$ converges superlinearly to θ^*.*

Proof. A Taylor expansion of $\log g(y;\theta)$ at the current iteration θ^k gives

$$\nabla \log g(y;\theta) = \nabla \log g(y;\theta^k) + \langle \nabla^2 \log g(y;\theta^k), \theta - \theta^k \rangle + R(|\theta - \theta^k|). \tag{43}$$

Take $\theta = \theta^*$ and $G_k = \nabla^2 \log g(y;\theta^k)$ in the last expansion, and obtain

$$\nabla \log g(y;\theta^*) = \nabla \log g(y;\theta^k) + \langle G_k, \theta^* - \theta^k \rangle + R(|\theta^* - \theta^k|). \tag{44}$$

Now recall that

$$\theta^{k+1} = \theta^k - (G_k - \lambda_k I_k)^{-1} \nabla \log g(y; \theta^k), \tag{45}$$

which gives

$$\nabla \log g(y; \theta^k) + G_k(\theta^{k+1} - \theta^k) - \lambda_k I_k(\theta^{k+1} - \theta^k) = 0. \tag{46}$$

Substract this last equation to (44), and obtain

$$G_k(\theta^* - \theta^{k+1}) + \lambda_k I_k(\theta^{k+1} - \theta^k) + r(\theta^* - \theta^k) = 0. \tag{47}$$

Then,

$$-G_k(\theta^* - \theta^{k+1}) + \lambda_k I_k(\theta^* - \theta^k) = r(\theta^* - \theta^k) + \lambda_k I_k(\theta^* - \theta^k). \tag{48}$$

Multiplying both sides by $(\theta^* - \theta^{k+1})^T$, we obtain

$$\begin{aligned}\langle \theta^* - \theta^{k+1}, -G_k(\theta^* - \theta^{k+1})\rangle + \lambda_k \langle \theta^* - \theta^{k+1}, I_k(\theta^* - \theta^{k+1})\rangle \\ = \langle \theta^* - \theta^{k+1}, r(\theta^* - \theta^k)\rangle + \lambda_k \langle \theta^* - \theta^{k+1}, I_k(\theta^* - \theta^k)\rangle\end{aligned}$$

Recall that $-\log g(y; \theta)$ is elliptic over $\mathcal{B}$, and thus, there exists a constant number $l_g(\mathcal{B})$ such that

$$\langle d, G_k g\rangle \geq l_g(\mathcal{B})|d|^2. \tag{49}$$

Combining this result with ellipticity of $I(\theta^k, .)$ over $\mathcal{B}$, and using the Cauchy Schwartz inequality, we easily obtain

$$(l_g(\mathcal{B}) + \lambda_k l(\mathcal{B}))|\theta^* - \theta^{k+1}|^2 \leq |\theta^* - \theta^{k+1}||r(\theta^* - \theta^k)| + \lambda_k |\theta^* - \theta^{k+1}| L(\mathcal{B})|\theta^* - \theta^k|. \tag{50}$$

Thus, we have

$$|\theta^* - \theta^{k+1}| \leq \frac{|r(\theta^* - \theta^k)| + \lambda_k L(\mathcal{B})|\theta^* - \theta^k|}{l_g(\mathcal{B}) + l(\mathcal{B})}. \tag{51}$$

Dividing both sides by $|\theta^* - \theta^k|$, we obtain

$$\frac{|\theta^* - \theta^{k+1}|}{|\theta^* - \theta^k|} \leq \frac{1}{l_g(\mathcal{B}) + l(\mathcal{B})}\Big(\frac{|r(\theta^* - \theta^k)|}{|\theta^* - \theta^k|} + \lambda L(\mathcal{B})|\Big). \tag{52}$$

Now, since $\lim_{k\to\infty} \frac{|r(\theta^* - \theta^k)|}{|\theta^* - \theta^k|} = 0$, and $\lim_{k\to\infty} \lambda_k = 0$, we may assume that θ^* is close enough to θ^*, and that λ_k is small enough, so that

$$\frac{|\theta^* - \theta^{k+1}|}{|\theta^* - \theta^k|} \leq \rho < 1. \tag{53}$$

Thus, θ^{k+1} also belongs to $\mathcal{B}$, and we may deduce recursively that $|\theta^{k+1} - \theta^*| \leq \rho|\theta^k - \theta^*|$ for all large enough k. Thus, the sequence $(\theta^k)_{k\in N}$ converges to θ^*. Moreover, since $\lim_{k\to\infty} \lambda_k = 0$ and $\lim_{k\to\infty} \frac{|r(\theta^* - \theta^k)|}{|\theta^* - \theta^k|} = 0$, we clearly obtain from (52) that

$$\lim_{k\to\infty} \frac{|\theta^* - \theta^{k+1}|}{|\theta^* - \theta^k|} = 0, \tag{54}$$

which completes the proof. $\square$

5.3 Approximation of the Relaxed EM algorithm

In the following, we establish approximation formulas for Relaxed EM iterations. Recall that the Kullback distance is then given by

$$I(\theta^k, \theta) = \mathsf{E}_{\theta^k}[\log \frac{k(y;\theta^k)}{k(y;\theta)}], \quad \text{with } k(x;\theta) = \frac{f(x;\theta)}{g(y;\theta)}.$$

Thus, we easily obtain

$$I(\theta^k, \theta) = \mathsf{E}_{\theta^k}[\log f(x;\theta^k)] - \log g(y;\theta^k) - \mathsf{E}_{\theta^k}[\log f(x;\theta)] + \log g(y;\theta).$$

Computing the hessian gives

$$\nabla^2 I(\theta^k, \theta^k) = \nabla^2 \log g(y;\theta^k) - \nabla^2 \mathsf{E}_{\theta^k}[\log f(x;\theta^k)].$$

Therefore, the second order approximate step of the Relaxed EM algorithm is given by

$$\theta^{k+1} = \operatorname{argmax}_{\theta \in R^p} \{ \ \langle \nabla \log g(y;\theta^k), \theta - \theta^k \rangle + \langle \theta - \theta^k, ((1-\lambda_k)G_k + \lambda_k F_k)(\theta - \theta^k) \rangle \ \}, \tag{55}$$

where $F_k = \nabla^2 \mathsf{E}_{\theta^k}[\log f(x;\theta^k)]$. Similarly to the study in the former section, we may deduce a quasi-second order method using BFGS updates of G_k, and the following approximated step

$$\theta^{k+1} = \theta^k - (W_k + \lambda_k W_k (F_k - W_k) W_k) \nabla \log g(y;\theta^k). \tag{56}$$

In the next section, these methods are applied to a problem with poisson data which occurs in PET image reconstruction.

6 Application to Poisson data and simulations results

In this section, we focus on a likelihood maximization problem with Poisson data which appears in emission tomography and we provide simulation results for the approximate Relaxed EM algorithm.

6.1 Presentation of the problem

The objective is to estimate the intensity vector $\theta = [\theta_1, \ldots, \theta_p]^T$ governing the number of gamma-ray emission $N = [N_1, \ldots, N_p]^T$ over an imaging volume of p pixels. The estimate of θ must be based on the observed data $Y = [Y_1, \ldots, Y_m]^T$. The components of N_i of N are independent Poisson distributed with rate parameters θ_i, and the components of Y_j of Y are independent Poisson distributed with rate parameters $\sum_{i=1}^p P_{j,i}\theta_i$, where $P_{j,i}$ is the transition probability corresponding to emissions from pixel i being detected at detector module j. The standard choice of data X for estimation of θ in EM context is the set $(N_{j,i})_{1 \le j \le m,\ 1 \le i \le p}$, where $N_{j,i}$ denotes the number of emissions in pixel i which are detected at detector j. The many-to-one application $X \mapsto h(X) = Y$ necessary to the contruction of the EM framework is

$$Y_j = \sum_{i=1}^{p} N_{j,i}, \quad 1 \leq j \leq m. \tag{57}$$

It is easily established that the variables $N_{j,i}$ are independent Poisson distributed with intensity $P_{j,i}\theta_i$ that the likelihood function is given by

$$\log g(y;\theta) = \sum_{j=1}^{m} \Big(\sum_{i=1}^{p} P_{j,i}\theta_i \Big) - y_j \log \Big(\sum_{i=1}^{p} P_{j,i}\theta_i \Big) + \log y_j! \tag{58}$$

and that the expectation step is (see [3])

$$\mathsf{E}_{\theta'}[\log f(x;\theta)] = \sum_{j=1}^{m}\sum_{i=1}^{p} \Big(\frac{y_j P_{j,i}\theta'_i}{\sum_{i=1}^{p} P_{j,i}\theta'_i} \log(P_{j,i}\theta_i) - P_{j,i}\theta_i \Big). \tag{59}$$

6.2 The approximate Relaxed EM algorithm for PET

In order to applied our approximate Relaxed EM methods, we need calculate the following quantities

$$\nabla \log g(y;\theta^k) = \Big[g_i \Big]_{1 \leq i \leq p} \qquad \text{with} \qquad \sum_{j=1}^{m} P_{j,i}\theta_i^k - \frac{y_j Pj,i}{\sum_{i=1}^{p} P_{j,i}\theta_i^k} \tag{60}$$

and

$$\nabla^2 \log g(y;\theta^k) = [G_{i,i'}^k]_{1 \leq i,i' \leq p} \tag{61}$$

with

$$\begin{cases} G_{i,i}^k = \sum_{j=1}^{m} P_{j,i} + \frac{y_j P^2 j,i}{(\sum_{i=1}^{p} P_{j,i}\theta_i^k)^2}, \\ \\ G_{i,i'}^k = \sum_{j=1}^{m} y_j \frac{P_{j,i}P_{j,i'}}{(\sum_{i=1}^{p} P_{j,i}\theta_i^k)^2} \qquad i \neq i' \end{cases} \tag{62}$$

We will also need the expression of the hessian of the expected quantity 59

$$\nabla^2 \mathsf{E}_{\theta^k}[\log f(x;\theta)] = \Big[F_{i,i'} \Big]_{1 \leq i,i' \leq p} \tag{63}$$

with

$$\begin{cases} F_{i,i} = -\frac{1}{\theta_i^{k\,2}} \sum_{j=1}^{m} \frac{y_j P_{j,i}\theta_j^k}{\sum_{i=1}^{p} P_{j,i}\theta_i^k}. \\ \\ F_{i,i'} = 0 \qquad i \neq i' \end{cases} \tag{64}$$

Thus, we have all the elements to compute the approximate Relaxed EM iterations. We easily deduce from 55 that the algorithm is defined by

$$\theta^{k+1} = \theta^k - ((1-\lambda_k)G_k + \lambda_k F_k)\nabla \log g(y;\theta^k), \tag{65}$$

with starting point $\theta^1 \in R^p$. We now present some simulation results

6.3 Simulation results

7 Conclusion

References

[1] J. F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. A. Sagastizabal. *Optimisation Numérique: Aspects théoriques et pratiques.* Springer-Verlag, 1997.

[2] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, 1994.

[3] P. J. Green. On the use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society Series B*, 52:443–452, 1990.

[4] A. N. Iusem, B. Svaiter, and M. Teboulle. Entropy-like proximal methods in convex programming. *Mathematics of Operations Research*, 19:790–814, 1994.

[5] A. N. Iusem and M. Teboulle. A primal-dual iterative algorithm for a maximum likelihood estimation problem. *Computational Statistics and Data Analysis*, 14(4):443–456, 1992.

[6] S. Kullback. *Information theory and statistics.* John Wiley and Sons, Inc., New York, 1959.

[7] B. Lemaire. The proximal point algorithm. pages 73–87, 1989. New methods in optimization and their industrial uses (Pau/Paris, 1987).

[8] X. L. Meng and D. B. Rubin. Maximum likelihood estimation via the ecm algorithm: a general framework. *Biometrika*, 80:267–278, 1993.

[9] J.-J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[10] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[11] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.