The EM algorithm for ill-posed integral equations: a convergence analysis

Elena Resmerita*, Heinz W. Engl* and Alfredo N. Iusem†

August 14, 2007

Abstract

The EM (Expectation-Maximization) algorithm is a convenient tool for approximating maximum likelihood estimators in situations when available data are incomplete, as it is the case for many inverse problems. Our focus here is on the continuous version of the EM algorithm for a Poisson model, which is known to perform unstably when applied to ill-posed integral equations. We interpret and analyse the EM algorithm as a regularization procedure: We show weak convergence of the iterates to a solution of the equation when exact data are considered. In the case of perturbed data, similar results are established by employing a stopping rule of discrepancy type under boundedness assumptions on the problem data.

1 Introduction

There is a large body of literature concerning the Expectation-Maximization (EM) algorithm, as introduced by Dempster, Laird and Rubin [5] in 1977, for approximating maximum likelihood estimators in problems with incomplete data. The importance of the methodolody resides in the (usually) simple form of the complete likelihood function that is to be maximized via EM, which sometimes even leads to explicit iterative formulas. The book [15] offers a comprehensive treatment for the finite dimensional case and points out that EM algorithms have a "tremendous potential for applications", being "the subject of numerous extensions" and "thousands of publications". In this paper, we refer only to a few of those publications, which are directly related to our work.

In the sequel, we briefly describe the EM methodology. Let $\mathbf{Y}$ be a random vector with a probability density function $p_{\mathbf{Y}}(\mathbf{y};\theta)$ depending on a parameter

*Johann Radon Institute for Computational and Applied Mathematics (RICAM),Austrian Academy of Sciences, Altenbergerstrasse 69,4040 Linz, Austria (elena.resmerita@ricam.oeaw.ac.at, heinz.engl@ricam.oeaw.ac.at)

†Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro, R.J., CEP 22460-320, Brazil (iusp@impa.br)

θ. The associated log-likelihood function is

$$L_{\mathbf{Y}}(\theta) = \log p_{\mathbf{Y}}(\mathbf{y}; \theta),$$

where $\mathbf{y}$ is a given realization of $\mathbf{Y}$. EM is an iterative method which provides approximations of a maximum log-likelihood estimator (MLE) for the parameter, given a realization $\mathbf{y}$ of $\mathbf{Y}$. By a MLE for θ given $\mathbf{y}$, one understands a parameter $\hat{\theta}$ such that

$$l_{\mathbf{Y}}(\hat{\theta}) = \max\, l_{\mathbf{Y}}(\theta).$$

The basic idea of the EM technique is to associate to the given incomplete-data problem, a complete-data problem for which the MLE problem can be conveniently solved. Denote by $\mathbf{X}$ the random vector corresponding to the complete data. The expectation (E) step consists of averaging the complete-data log likelihood over its conditional distribution, given the observed data (by using the current iterate for the unknown parameter). That is, given a realization $\mathbf{y}$ of $\mathbf{Y}$ and an initial guess θ^0, compute

$$Q(\theta, \theta^k) = E[\log p_{\mathbf{X}}(\mathbf{x}; \theta)|\mathbf{y}; \theta^k], \;\; k \geq 0.$$

In the maximization (M) step, one maximizes the conditional expectation:

$$\theta^{k+1} \in argmax Q(\theta, \theta^k).$$

In the particular situation when the algorithm involves only independently distributed Poisson variables, the algorithm reduces to a simple closed formula. More precisely, this EM has the following expression:

$$x_{k+1}^j = x_k^j \sum_{i=1}^{n} \frac{a_{ij} y_i}{\sum_{l=1}^{m} a_{il} x_k^l}, \;\; j = 1, m, \tag{1}$$

where all the involved variables are nonnegative. An important application of this algorithm is in processing images by Positron Emission Tomography (PET) - see [22], [13], [25]. PET works by first introducing a radioactive substance into the region of interest of the body; following some reactions, positrons are emitted and recorded by means of a ring of detectors placed around the body. This process is frequently used in medical diagnosis to identify tumors and various diseases. The notation in (1) has the following meaning in this context: The vector x provides information about the number of positron emissions from the body section whose structure is to be recovered, the matrix $A = (a_{ij})$ contains the probabilities with which the emissions are detected, while y counts the detected emissions. The log-likelihood function is, in this case, the so-called "Kullback-Leibler divergence", namely

$$d(y, Ax) := \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{(Ax)_i} - y_i + (Ax)_i \right].$$

The reader is refered to [10] for more details.

Convergence of algorithm (1) to a maximum likelihood estimator was shown in [4], [25], [11] in various situations. However, it has been observed through numerical experiments that the algorithm may be unstable (see, [25], [23]); e.g., as noticed in [23], the EM (1) outlines the shape and intensity of the objects in early iterations, while the later iterations roughen the image. This has been the motivation for several authors (see, e.g., [26], [2]) to propose stopping rules for this procedure, mostly in statistical settings. The main cause for the unstable behaviour of the algorithm is the inverse nature and hence, ill-posedness of the considered problems. Since ill-posedness is an infinite dimensional phenomenon, an analysis of the algorithm in infinite dimensional spaces in the framework of regularization can be a good starting point for dealing with the instability.

The continuous version of procedure (1), that is,

$$x_{k+1}(t) = x_k(t) \int_\Sigma \frac{a(s,t)y(s)}{(Ax_k)(s)}\, ds, \tag{2}$$

was proposed as early as 1983 in [12] as a method for solving Fredholm integral equations of the first kind:

$$Ax = y, \tag{3}$$

with

$$(Ax)(s) = \int_\Omega a(s,t)x(t)\, dt. \tag{4}$$

However, it was [17] that indicated the connection between (2) and its famous discrete counterpart, and analysed the former along the same lines as in the finite dimensional case. One could think of the function

$$f(x) := \int_\Sigma \left[y(s) \log \frac{y(s)}{Ax(s)} - y(s) + Ax(s) \right] ds \tag{5}$$

as of the log-likelihood function for infinitely many samples. The existing results concerning algorithm (2) mainly state that the fixed points of the algorithm are minimizers of the function f (in particular, solutions of the operator equation), that the function f decreases along the iterations and that the Kullback-Leibler distance between the solution and the iterates decreases, too (cf. [17], [18], [19]) - see Section 3 for details.

To the best of our knowledge, there is no result in the EM literature showing convergence of the algorithm for PET in infinite dimensional spaces, but only the partial results mentioned above. There is no hope that the algorithm converges strongly in Lebesgue spaces if the function f does not have minimizers (cf. [19]). Algorithm stabilization has been attempted by means of, e.g., smoothing steps (see [23], [14], [6]) or penalized maximum likelihood ([7]). Both approaches have drawbacks, as they depend on smoothing operators and penalty functions plus the right choice of the penalization parameter, respectively. An early termination of the iterative procedure, according to an appropriate stopping rule, will be the key to overcome the instability issue. The reader may consult [8] for an overview over stopping rules for iterative regularization methods.

We emphasize that the EM algorithm is considered here in a deterministic framework, rather than in a statistical (stochastic) one.

The paper is organized as follows: Section 2 presents notation, main assumptions and recalls several useful results. A brief review of the existing work regarding algorithm (2) will be presented in Section 3. Then we prove strong convergence (in Lebesgue spaces) of the images of the iterates to the exact data; moreover, under the additional assumptions that the initial point x_0 is a bounded function and the equation has a bounded solution, we show weak convergence of the iterates to the solution, in Section 4. Also, we consider in (2) perturbed data y^δ instead of exact data y, with

$$\|y^\delta - y\|_1 \leq \delta, \;\; \delta > 0. \tag{6}$$

In Section 5 we establish several intermediate results for noisy data, prove monotonicity of the residual and show that the iterates get closer to the solution with respect to the Kullback-Leibler divergence until the residual reaches a certain threshold. Based on these, we show in Section 6 that the perturbed EM algorithm with a rule for stopping at a certain iteration is indeed a regularization procedure for the ill-posed equation (3). That is, the images of the iterates corresponding to the stopping index (depending on δ and y^δ) converge strongly to the exact data, and the iterates converge weakly to the solution as the noise level δ tends to zero. The assumptions under which these results have been obtained are not as restrictive as they might seem at first sight - see the discussion at Remark 6.1.

2 Notation and assumptions

The sets Ω, $\Sigma \subset \mathbb{R}^d$, $d \geq 1$, are compact. Let Δ stand for the set

$$\Delta = \{u \in L^1(\Sigma) : u \geq 0, \int_\Sigma u(s)\, ds = 1\}. \tag{7}$$

For a function g defined on a space X, we denote the domain of g by $dom\, g = \{x \in X : g(x) < +\infty\}$. Throughout this paper we assume that the kernel a is a positive and measurable function such that the operator $A : L^1(\Omega) \to L^1(\Sigma)$ defined by (4) is continuous.

The main assumptions of the present work are:

(A1) *The kernel a satisfies the following normalization condition*:
For almost any $t \in \Omega$,

$$\int_\Sigma a(s,t)\, ds = 1; \tag{8}$$

(A2) *The kernel a is bounded and bounded away from zero, in the following sense*:

There exist $m, M > 0$ such that

$$m \leq a(s,t) \leq M, \;\; \text{a.e.} \;\; \text{on} \;\; \Sigma \times \Omega. \tag{9}$$

(A3) *The exact data y in (3) belong to the set Δ.*

(A4) *There is a nonnegative solution z of equation (3), and it does not vanish a.e.*

Moreover, we assume that there exists $M' > 0$ such that

$$y(s) \leq M', \quad \text{a.e.} \quad \text{on } \Sigma. \tag{10}$$

Note that the right hand-side inequality in (9) ensures that A is a compact operator at least from $L^1(\Omega)$ to $L^1(\Sigma)$ and therefore, it is also continuous in this setting.

The Kullback-Leibler (KL) divergence or distance, denoted below by d, is the functional defined by

$$d(v,u) = \int \left[v(t) \ln \frac{v(t)}{u(t)} - v(t) + u(t) \right] dt, \tag{11}$$

whenever it is finite*. The domain of integration will be either Ω or Σ, depending on the context. Note that $d(u,v) \geq 0$, for any (u,v) in $dom\, d$, and

$$d(u,v) = 0 \text{ if and only if } u = v. \tag{12}$$

The KL-functional will play a crucial role in the ensuing analysis. Therefore, we recall some of its algebraic and topological properties (see, e.g., [21]):

Lemma 2.1 i) For any $(u,v) \in dom\, d$, it holds that

$$\|v-u\|_1^2 \leq \left(\frac{2}{3}\|v\|_1 + \frac{4}{3}\|u\|_1 \right) d(v,u). \tag{13}$$

ii) The function $(v,u) \mapsto d(v,u)$ is convex and thus, the same holds for the function $(v,x) \mapsto d(v,Ax)$.

Corollary 2.2 If $\{u_\lambda\}_\lambda$, $\{v_\lambda\}_\lambda$ are sequences in $L^1(\Omega)$ such that one of them is bounded, then

$$\lim_\lambda d(v_\lambda, u_\lambda) = 0 \quad \Longrightarrow \quad \lim_\lambda \|v_\lambda - u_\lambda\|_1 = 0. \tag{14}$$

3 A brief review of the properties of the EM algorithm

From now on, we refer to the EM algorithm in infinite dimensional spaces only. Let $x_0 \in \Delta$ such that $x_0 > 0$, a.e. Denote by

$$P(x) := x \int_\Sigma \frac{a(s,\cdot) y(s)}{(Ax)(s)}\, ds \tag{15}$$

*We use the convention $0 \log 0 = 0$.

and let

$$\lambda_x(s,t) := \frac{a(s,t)x(t)}{(Ax)(s)}.$$

Thus, we have, for any $t \in \Omega$,

$$P(x)(t) = \int_\Sigma y(s)\lambda_x(s,t)\,ds$$

and then (2) can be written as

$$x_{k+1} = P(x_k). \tag{16}$$

Observe that

$$\int_\Omega \lambda_x(s,t)\,dt = 1, \ \ \forall s \in \Sigma. \tag{17}$$

When solutions exist for equation (3), one approach to find them is, due to (12), minimizing the convex function f defined by (5). It has been proven that the minimizers of the function f are fixed points of the corresponding operator P. This and other properties described below have been shown in [18] under continuity and positivity assumptions on the kernel a and the data y, as well as in [6] for a slightly different algorithm.

Proposition 3.1 Let (A1) and (A3) be satisfied and consider $x \in \Delta \cap dom\, f$. Then the following assertions hold:

i) $P(x) \in \Delta \cap dom\, f$ and

$$d(P(x), x) \leq f(x) - f(P(x)). \tag{18}$$

ii) If z is a minimizer of the function f such that $d(z,x) < \infty$, then one has $d(z, P(x)) < \infty$ and

$$f(x) - f(z) \leq d(z,x) - d(z, P(x)). \tag{19}$$

Proof: See [18] or [6]. For a different proof technique, see the forthcoming proof of Proposition 5.2, with $y^\delta = y$. □

Corollary 3.2 Every minimizer of the function f is a fixed point of the operator P defined by (15).

Proof: It follows immediately from Proposition 3.1(i) and (12). □

Proposition 3.3 Let (A1) and (A3) be satisfied and consider a minimizer z of the function f. Then the EM algorithm (2) has the following properties, for any $k \in \mathbb{N}$:

$$d(x_{k+1}, x_k) \leq f(x_k) - f(x_{k+1}), \tag{20}$$

$$f(x_k) - f(z) \leq d(z, x_k) - d(z, x_{k+1}). \tag{21}$$

Therefore, the sequences $\{d(z,x_k)\}_{k\in\mathbb{N}}$, $\{f(x_k)\}_{k\in\mathbb{N}}$ are nonincreasing and, moreover, $\lim_{k\to\infty} f(x_k) = f(z)$ and $\lim_{k\to\infty} d(x_{k+1}, x_k) = 0$. If, in addition, $d(z, x_0) < \infty$, then $d(z, x_k) < \infty$, $\forall k \in \mathbb{N}$.

Proof: Inequalities (20) and (21) result by taking $x = x_k$ in the inequalities shown in Proposition 3.1. If $d(z, x_0) < \infty$, then inequality (21) ensures that $d(z, x_k) < \infty$, $\forall k \in \mathbb{N}$. □

This further yields $\lim_{k\to\infty} \|Ax_k - y\|_1 = 0$ and $\lim_{k\to\infty} \|x_{k+1} - x_k\|_1 = 0$ (cf. Corollary 2.2). Consequently, the sequence $\{d(z, x_k)\}_{k\in\mathbb{N}}$ converges to some nonnegative number. In the finite dimensional case it has been proven, based on the boundedness of $\{\|x_k\|_1\}_{k\in\mathbb{N}}$, that this number is zero, implying convergence of $\{x_k\}_{k\in\mathbb{N}}$ to z (see, e.g., [11]). However, this has not been the case in infinite dimensional spaces.

4 Convergence of the EM algorithm for exact data

We show in the sequel that the iterates x_k produced by (2) converge to a solution z of equation (3) with respect to weak topologies on Lebesgue spaces.

Let us start with a consequence of the assumptions described in Section 2.

Proposition 4.1 If (A2) holds, then for any $x \in \Delta$,

$$m \leq (Ax)(s) \leq M, \text{ a.e. on } \Sigma. \tag{22}$$

Proof: Multiply (9) by $x(t)$ for any $t \in \Omega$, then integrate over Ω. □

We shall also need the following results:

Lemma 4.2 If $\{h_n\}_{n\in\mathbb{N}}$ is a sequence of Lebesgue integrable functions on a space of finite measure, which is uniformly bounded and such that

$$\lim_{n\to\infty} \|h_n - h\|_1 = 0,$$

then for any $p \in [1, +\infty)$,

$$\lim_{n\to\infty} \|h_n - h\|_p = 0.$$

Proof: Let $c > 0$ be such that $\|h_n\|_\infty \leq c$, for any $n \in \mathbb{N}$. Since $h_n \xrightarrow{L^1} h$ as $n \to \infty$, we have that $h_n \xrightarrow{a.e.} h$. This and the hypotheses imply that h is a bounded function. Fix $p > 1$. Then the following inequality holds for every t:

$$|h_n(t) - h(t)|^p \leq 2^{p-1}|h_n(t) - h(t)| \cdot (|h_n(t)| + |h(t)|)^{p-1}.$$

By integrating over the set of t, we obtain the result. □

Proposition 4.3 For $0 \leq x, u \in L^\infty(S, d\mu)$ and $2 \leq p < +\infty$,

$$d(x, u) \geq \frac{1}{p(p-1)} (\max\{\|x\|_\infty, \|u\|_\infty\})^{1-p} \|x - u\|_p^p. \tag{23}$$

Proof: See Proposition 2.3 in [3]. □

Now we state the convergence result for the EM algorithm.

Theorem 4.4 Let assumptions (A1)-(A4) and inequality (10) be satisfied. If the initial point $x_0 \in \Delta$ is such that $d(z, x_0) < \infty$, then

i)
$$\lim_{k\to\infty} \|Ax_k - y\|_p = 0, \tag{24}$$
for any $p \in [1, +\infty)$.

ii) If the initial point x_0 is a bounded function and equation (3) has a bounded solution z, then the iterates x_k produced by (2) have a subsequence that converges to a solution of the equation in the weak* topology of $L^\infty(\Omega)$ and therefore, in the weak topology of $L^p(\Omega)$ for any $p \in [1, +\infty)$. If, in addition, the solution is unique, then the whole sequence converges to the solution z in the same topologies.

Proof:

i) Equality (24) with $p = 1$ follows immediately - see the discussion following Proposition 3.3. Then Proposition 4.1 and Lemma 4.2 imply (24), for any $1 \le p < +\infty$.

ii) We show by induction that any x_k, $k \in \mathbb{N}$, is a bounded function. Let $k = 0$. Inequalities (10) and (22) imply $\lambda_{x_0}(s,t) \le \frac{M'M}{m}\mu(\Sigma)$, a.e. on Ω, and then boundedness of x_1 a.e. is ensured. One can similarly deal with the k-th induction step, showing that $x_k(t) \le c_k$ a.e., where c_k are positive numbers. Consequently, $\{x_k\}_{k\in\mathbb{N}}$ is contained in $L^\infty(\Omega)$. We claim that $\{x_k\}_{k\in\mathbb{N}}$ is bounded with respect to the $L^\infty(\Omega)$ norm. To this end, suppose by contradiction that there exists a subsequence $\{x_n\}_{n\in\mathbb{N}}$ such that $\|x_n\|_\infty \to +\infty$. Then $\max\{\|x_n\|_\infty, \|z\|_\infty\} = \|x_n\|_\infty$ for n sufficiently large, which together with Proposition 4.3 yields

$$\begin{aligned} d(z, x_n) &\ge \frac{1}{p(p-1)}(\max\{\|x_n\|_\infty, \|z\|_\infty\})^{1-p}\|x_n - z\|_p^p \qquad (25)\\ &= \frac{1}{p(p-1)}\frac{\|x_n - z\|_p^p}{\|x_n\|_\infty^{p-1}} \\ &= \frac{\|x_n\|_\infty}{p(p-1)}\frac{\|x_n - z\|_p^p}{\|x_n\|_\infty^{p}} \end{aligned}$$

for some $p \in [2, +\infty)$. Observe that the quantity $\frac{\|x_n - z\|_p}{\|x_n\|_\infty}$ is bounded since

$$\begin{aligned} \frac{\|x_n - z\|_p}{\|x_n\|_\infty} &\le \frac{\|x_n\|_p}{\|x_n\|_\infty} + \frac{\|z\|_p}{\|x_n\|_\infty} \\ &\le \mu(\Omega)^{1/p} + \frac{\|z\|_p}{\|x_n\|_\infty}. \end{aligned}$$

This and the observation that $d(z, x_n) \leq d(z, x_0)$ for any $n \in \mathbb{N}$ imply that the last inequality in (25) cannot hold. Thus, our claim is proved. Hence, by Alaoglu-Bourbaki Theorem, [9, p. 70], there is a subsequence $\{x_j\}_{j\in\mathbb{N}}$ which converges in the weak* topology of $L^\infty(\Omega)$ to some $u \in L^\infty(\Omega)$. This means that $\langle x_j - u, \varphi\rangle \to 0$ for any $\varphi \in L^1(\Omega)$, as $j \to \infty$. In particular, this holds also for any $\varphi \in L^\infty(\Omega)$. That is, the sequence $\{x_j\}_{j\in\mathbb{N}}$ converges weakly in $L^1(\Omega)$ to u. Since A is continuous and linear, it is weakly continuous. Hence, it follows that $\{Ax_j\}_{j\in\mathbb{N}}$ converges weakly in $L^1(\Sigma)$ to Au as $j \to \infty$. This sequence converges also strongly, and then, weakly to y in $L^1(\Sigma)$ (see (24)). Thus, it follows that $Au = y$. Note that weak* convergence in $L^\infty(\Omega)$ implies weak convergence in $L^p(\Omega)$, for any $p \in [1, +\infty)$ (this follows immediately from the definition, since $\mu(\Omega) < +\infty$). Thus, the subsequence $\{x_j\}_{j\in\mathbb{N}}$ converges weakly in these spaces. If the solution z is unique, then the whole sequence converges to z with respect to the above mentioned topologies since, otherwise, every divergent subsequence would have a subsequence which would converge to z.

□

Remarks.

- The assumption regarding $d(z, x_0)$ is reasonable from a practical point of view. Indeed, the initial point can be chosen as a positive constant function (this is usually the case in PET - see [23]), while the solution z should be such that $\int_\Omega z(t) \log z(t)\, dt < \infty$.
- Although $\{Ax_k\}_{k\in\mathbb{N}}$ converges to y in any $L^p(\Omega)$ with $p \in [1, +\infty)$ (see (24)), this does not necessarily imply strong convergence in $L^\infty(\Omega)$; one would have this if (24) held uniformly with respect to p.
- We would like to point out implications of the weak* convergence of the algorithm in $L^\infty(\Omega)$, which has been shown above: In case of unique solution and exact data, we obtain that

$$\int_\Omega x_k \varphi\, d\mu \to \int_\Omega z\varphi\, d\mu, \text{ as } k \to \infty,$$

for any $\varphi \in L^1(\Omega)$. This is useful, e.g., in situations when one is interested not in the solution z, but rather in some linear functional of the solution $\langle \varphi, z\rangle$ (see [1]). Another consequence is that z belongs to Δ, which is not surprising.

5 The case of noisy data

We assume that only noisy data y^δ is available for the considered ill-posed equation. More precisely,

(A5) *The perturbed data y^δ belong to $\Delta \cap L^\infty(\Sigma)$ and satify (6).*

Our aim is to show several properties of the following

Iterative procedure:

Choose $x_0 \in \Delta$ to be positive a.e. and let, for any integer $k \geq 0$,

$$x_{k+1}^{\delta} = P^{\delta}(x_k^{\delta}), \tag{26}$$

where $x_0^{\delta} := x_0$ and

$$P^{\delta}(x) := x \int_{\Sigma} \frac{a(s,\cdot)y^{\delta}(s)}{(Ax)(s)}\, ds. \tag{27}$$

To this end, we point out several inequalities related to the operators P and P^{δ}. We shall need the following version of Jensen's inequality (see [20]):

Lemma 5.1 Let $\tilde{\mu}$ be a positive measure on a σ-algebra in the set Σ with $\tilde{\mu}(\Sigma) = 1$. If f_1, f_2 are real functions in $L^1(\Sigma, \tilde{\mu})$, $a < f_i(x) < b$ for all $x \in \Sigma$, with $a, b \in \bar{\mathbb{R}}$, $i = 1, 2$, and if φ is convex on $(a, b) \times (a, b)$, then

$$\varphi\left(\int_{\Sigma} f_1 d\tilde{\mu}, \int_{\Sigma} f_2 d\tilde{\mu}\right) \leq \int_{\Sigma} \varphi \circ (f_1, f_2) d\tilde{\mu}. \tag{28}$$

In fact, we shall use the case when $\varphi : (0, +\infty) \times (0, +\infty) \to (0, +\infty)$ is the following function of two variables, which is jointly convex (see Lemma 2.1):

$$\varphi(s_1, s_2) = s_1 \ln \frac{s_1}{s_2} - s_1 + s_2.$$

If $\mu(\Sigma)$ does not equal one but it is still finite, inequality (28) above still holds when the integral is taken with respect to μ since we can define $\tilde{\mu}(S) := \frac{\mu(S)}{\mu(\Sigma)}$ for any $S \subset \Sigma$ and take advantage of the fact that φ is positively homogeneous.

For fixed $\delta > 0$, let f^{δ} denote the function

$$f^{\delta}(x) := d(y^{\delta}, Ax). \tag{29}$$

Proposition 5.2 Let assumptions (A1) and (A3) be satisfied. Then, for any $u, v, w \in \Delta \cap dom\, f$, the following inequalities hold:

$$d(P(w), P^{\delta}(v)) \leq d(y, y^{\delta}) + d(P(w), v) - d(P(w), w) + f(w) - f(v), \tag{30}$$

$$d(P^{\delta}(u), P^{\delta}(v)) \leq d(P^{\delta}(u), v) - d(P^{\delta}(u), u) + f^{\delta}(u) - f^{\delta}(v). \tag{31}$$

Proof: We prove only the first inequality, since the second one results in a similar manner. To this end, we adapt to our framework a proof idea used in [11] for establishing a similar inequality for the exact data case in a finite

dimensional setting. By Jensen's inequality (28), Fubini's Theorem and (17), we have

$$\begin{aligned}
d(P(w), P^\delta(v)) &= \int_\Omega \varphi\left(\int_\Sigma y(s)\lambda_w(s,t)\,ds, \int_\Sigma y^\delta(s)\lambda_v(s,t)\,ds\right)\,dt \\
&\le \int_\Omega \int_\Sigma \varphi\left(y(s)\lambda_w(s,t), y^\delta(s)\lambda_v(s,t)\right)\,ds\,dt \\
&= \int_\Omega \int_\Sigma y(s)\lambda_w(s,t)\ln\frac{y(s)\lambda_w(s,t)}{y^\delta(s)\lambda_v(s,t)}\,ds\,dt \\
&= \int_\Sigma y(s)\ln\frac{y(s)}{y^\delta(s)}\int_\Omega \lambda_w(s,t)\,dt\,ds \\
&+ \int_\Omega \int_\Sigma y(s)\lambda_w(s,t)\ln\frac{w(t)}{v(t)}\,ds\,dt \\
&+ \int_\Sigma y(s)\ln\frac{Av(s)}{Aw(s)}\int_\Omega \lambda_w(s,t)\,dt\,ds \\
&= d(y,y^\delta) + [d(P(w),v) - d(P(w),w)] + [f(w) - f(v)].
\end{aligned}$$

□

Proposition 5.3 Let assumptions (A1), (A3) and (A4) be satisfied. Then for any $\delta > 0$ and $k \in \mathbb{N}$, it holds that

$$f(x_k^\delta) - d(y, y^\delta) \le d(z, x_k^\delta) - d(z, x_{k+1}^\delta), \tag{32}$$

$$f^\delta(x_k^\delta) + \int_\Sigma \left[y(s) - y^\delta(s)\right]\ln\frac{y^\delta(s)}{(Ax_k^\delta)(s)}\,ds \le d(z, x_k^\delta) - d(z, x_{k+1}^\delta), \tag{33}$$

$$d(x_{k+1}^\delta, x_k^\delta) + d(x_{k+1}^\delta, x_{k+2}^\delta) \le f^\delta(x_k^\delta) - f^\delta(x_{k+1}^\delta). \tag{34}$$

Proof: Since the solution z is a minimizer of the function f, it follows that $P(z) = z$ (cf. Corollary 3.2). Letting $w = z$ and $v = x_k^\delta$ in (30) proves inequality (32). A simple calculation yields (33) from (32). Letting $u = x_k^\delta$ and $v = x_{k+1}^\delta$ in (31) implies (34). □

An immediate consequence of inequality (34) is that $\{f^\delta(x_k^\delta)\}_{k\in\mathbb{N}}$ is non-increasing in k for any fixed $\delta > 0$. However, a similar property cannot be expected to hold for $d(z, x_k^\delta)$. As discussed, e.g., in [8, Chapter 6], an iterative method for ill-posed problems shows the following typical behavior: The (metric) distance between the iterates and the solution has an initial decay and then increases. Thus, stopping the algorithm at a "good" index would ensure that the iterative method does regularize the problem, i.e., it provides stable approximations of the true solution. A famous and widely used stopping rule is Morozov's "discrepancy principle" ([16]).

6 A stopping rule for the EM algorithm

In this section, we provide a discrepancy type stopping rule for the EM algorithm (26) and establish existence of the corresponding stopping index $k_*(\delta)$. Also, we prove strong convergence of the images $\{Ax^\delta_{k_*(\delta)}\}_{\delta>0}$ to the exact data y, as $\delta \to 0^+$, with respect to any L^p norm, with $p \in [1, +\infty)$, along with the property that the iterates get closer to the solution with respect to the Kullback-Leibler divergence until the stopping index $k_*(\delta)$ is reached. In addition, if the initial point x_0 is a bounded function and the equation has a bounded solution, then weak convergence on subsequences of the iterates $x^\delta_{k_*(\delta)}$, as $\delta \to 0_+$, to a solution of the equation is obtained. This actually shows that the procedure (39) together with the rule (40) is a regularization method.

In addition to assumptions (A1)-(A5), we consider the following:

The perturbed data y^δ are bounded and bounded away from zero, i.e.,

$$m_1 \leq y^\delta(s) \leq M_1, \quad \text{a.e.} \quad \text{on } \Sigma, \tag{35}$$

uniformly for $\delta > 0$.

Remark 6.1 In most practical applications one can assume that y^δ and $a(s,t)$ are bounded from above. We can also assume that y^δ is bounded away from zero. Indeed, if $y^\delta(s)$ vanishes in $\Sigma_0 \subset \Sigma$, then the solution $z(t)$ vanishes in $\Omega_0 := \{t \in \Omega : \mu(\{s \in \Sigma_0 : a(s,t) > 0\}) > 0\}$, and we redefine the problem with the sets $\Omega \setminus \Omega_0$, $\Sigma \setminus \Sigma_0$, instead of Ω, Σ respectively, so that $y^\delta(s) > 0$ for all s in the new domain. We can further suppose that there exists $m_1 > 0$ such that $y(s) \geq m_1$ for all s, if we have in mind practical situations where sufficient data can be acquired.

The main problem lies in the assumption that $a(s,t) > m$ for almost all s, t, which fails in some real-world applications (e.g., positron emission tomography). However, the ensuing analyis does not exclude the more realistic case when the kernel fulfills simultaneously the following conditions:

i) $a(s,t)$ satisfies (9) on a set $\Sigma_0 \times \Omega_0 \subseteq \Sigma \times \Omega$ with $\mu(\Sigma_0 \times \Omega_0) > 0$;

ii) $a(s,t) = 0$ on a set $\Sigma_1 \times \Omega_1 \subsetneq \Sigma \times \Omega$ of positive Lebesgue measure.

iii) a is arbitrary on a set of measure zero.

It should be noted that conditions on the kernel and on the (exact) data similar to (A2) and (10), (35) are frequently employed when analyzing the EM in infinite dimensional spaces (see [17], [18], [19]), and might hold in other applications (see e.g. [24]).

Lemma 6.2 If (A2) and (35) hold, then, for all $\delta > 0$ and all $k \in \mathbb{N}$, we have that y^δ and Ax^δ_k, as well as $\ln y^\delta$ and $\ln Ax^\delta_k$ belong to $L^\infty(\Sigma)$.

Proof: follows from (35) and Proposition 4.1. □

We show below that the iterates keep approaching the solution so long as the residual $d(y^\delta, Ax^\delta_k)$ exceeds a certain threshold.

Theorem 6.3 Fix $\delta > 0$. If assumptions (A1)-(A5) and inequality (35) are satisfied, then the iterate x_{k+1}^{δ} is a better approximation of z than x_k^{δ} with respect to the Kullback-Leibler divergence d, that is,

$$d(z, x_{k+1}^{\delta}) \leq d(z, x_k^{\delta}), \ \ \forall k \in \mathbb{N}, \tag{36}$$

provided that k is such that

$$f^{\delta}(x_k^{\delta}) \geq \delta \max\left\{ \ln\left|\frac{M_1}{m}\right|, \ln\left|\frac{M}{m_1}\right| \right\}. \tag{37}$$

Proof: By applying the Cauchy-Schwarz inequality in (33), and using then inequalities (6), (35) and (22), we get

$$\begin{aligned} d(z, x_k^{\delta}) - d(z, x_{k+1}^{\delta}) &\geq f^{\delta}(x_k^{\delta}) - \|y - y^{\delta}\|_1 \left\| \ln \frac{y^{\delta}}{Ax_k^{\delta}} \right\|_{\infty} \\ &\geq f^{\delta}(x_k^{\delta}) - \delta \max\left\{ \ln\left|\frac{M_1}{m}\right|, \ln\left|\frac{M}{m_1}\right| \right\}, \end{aligned} \tag{38}$$

for any $\delta > 0$ and $k \in \mathbb{N}$. □

Therefore, we are motivated to propose the following

Stopping rule:

Let $x_0 \in \Delta$ be positive and

$$x_{k+1}^{\delta}(t) := x_k^{\delta}(t) \int_{\Sigma} \frac{a(s,t) y^{\delta}(s)}{(Ax_k^{\delta})(s)} \, ds, \ \ a.e. \tag{39}$$

For this EM algorithm, choose the stopping index as

$$k_*(\delta) = \min\left\{ k \in \mathbb{N} : f^{\delta}(x_k^{\delta}) \leq \tau \delta \max\left\{ \ln\left|\frac{M_1}{m}\right|, \ln\left|\frac{M}{m_1}\right| \right\} \right\}, \tag{40}$$

for some fixed $\tau > 1$, where M, m, M_1, m_1 are constants for which (A2) and (35) hold.

We show next that the index $k_*(\delta)$ given by (40) does exist and prove that the EM algorithm (39) is indeed a regularization method.

Theorem 6.4 Let assumptions (A1)-(A5) and inequality (35) hold, let $x_0 \in \Delta$ such that $d(z, x_0) < \infty$ and choose $k_*(\delta)$ according to the stopping rule (40). Then

i) The following inequality holds:

$$k_*(\delta)\tau\delta\gamma \leq k_*(\delta) f^{\delta}(x_{k_*(\delta)-1}^{\delta}) \leq d(z, x_0) + k_*(\delta)\delta\gamma, \ \ \forall \delta > 0, \tag{41}$$

where

$$\gamma = \max\left\{ \ln\left|\frac{M_1}{m}\right|, \ln\left|\frac{M}{m_1}\right| \right\}.$$

ii) The stopping index $k_*(\delta)$ is finite with $k_*(\delta) = O\left(\delta^{-1}\right)$ and

$$\lim_{\delta \to 0^+} \|Ax^{\delta}_{k_*(\delta)} - y\|_p = 0, \tag{42}$$

for any $p \in [1, +\infty)$.

iii) If the initial point x_0 is a bounded function and equation (3) has a bounded solution z, then a subsequence of the iterates $x^{\delta}_{k_*(\delta)}$ converges to a solution of the equation in the weak* topology of $L^\infty(\Omega)$, and therefore, in the weak topology of $L^p(\Omega)$ for any $p \in [1, +\infty)$, when $\delta \to 0^+$. If, in addition, the solution is unique, then the whole sequence converges to the solution z in the same topologies.

Proof:

i) Inequality (38) implies

$$f^{\delta}(x^{\delta}_j) + d(z, x^{\delta}_{j+1}) \leq d(z, x^{\delta}_j) + \delta\gamma,$$

for any $0 \leq j \leq k_*(\delta) - 1$. By summing up these inequalities and taking into account monotonicity of $f^{\delta}(x^{\delta}_j)$ with respect to j (cf. (34)) and (40), we obtain (41).

ii) From inequality (41) we deduce that the stopping index $k_*(\delta)$ is finite:

$$k_*(\delta) \leq \frac{d(z, x_0)}{\delta(\tau - 1)\gamma}. \tag{43}$$

Since, by (40),

$$d(y^{\delta}, Ax^{\delta}_{k_*(\delta)}) \leq \tau\delta\gamma, \tag{44}$$

it results that

$$\lim_{\delta \to 0^+} d(y^{\delta}, Ax^{\delta}_{k_*(\delta)}) = 0.$$

Consequently, equality (42) with $p = 1$ follows immediately due to (6) and Corollary 2.2. As of $1 \leq p < +\infty$, Corollary 6.2 and Lemma 4.2 imply $\lim_{\delta \to 0^+} \|Ax^{\delta}_{k_*(\delta)} - y\|_p = 0$.

iii) The proof can be done analogously to the one for the second part of Theorem 4.4; the key fact is showing that $\{x^{\delta}_{k_*(\delta)}\}_{\delta > 0}$ is bounded with respect to the $L^\infty(\Omega)$ norm, uniformly in δ. Also, one has to take into account that the net $\{x^{\delta}_{k_*(\delta)}\}_{\delta > 0}$ is a relatively weak*-sequentially compact set in $L^\infty(\Omega)$ (cf. [27, p. 72]), since $L^\infty(\Omega)$ is the dual of the separable Banach space $L^1(\Omega)$.

□

7 Conclusions

We have shown that the EM algorithm (39) converges weakly in Lebesgue spaces and, together with the discrepancy type rule (6), regularizes the ill-posed equation (3). It seems that these are the first convergence and regularization results to date for this procedure in infinite dimensional spaces.

So far, our theory allows only kernels which are bounded away from zero on a set of nonzero Lebesgue measure. We hope to be able to relax the conditions under which the results hold.

Strong convergence and convergence rates for the algorithm are certainly further challenges. Moreover, we are interested in designing an EM algorithm for nonlinear equations for which the analysis made in this work may be carried out under suitable assumptions on the involved operators.

8 Acknowledgments

The first author thanks Anca Croitoru (A.I. Cuza University, Iasi) for helpful discussions.

References

[1] Anderssen R S 1986 The linear functional strategy for improperly posed problems *Inverse Problems, Oberwolfach* (Basel: Birkhäuser) 11-30

[2] Bissantz N, Mair B and Munk A 2006 A multi-scale stopping criterion for MLEM reconstructions in PET *preprint*

[3] Borwein J M and Lewis A S 1991 Convergence of best entropy estimates *SIAM J. Optim.* 1(2) 191-205

[4] Csiszar I and Tusnady G 1984 Information geometry and alternating minimization procedures *Statistics and Decisions* Supplemental Issue Number 1 205-237

[5] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *Journal of the Royal Statistical Society* Series B 39(1) 1-38

[6] Eggermont P P B 1999 Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind *Applied Mathematics and Optimization* 39 75-91

[7] Eggermont P P B and LaRiccia V 1996 Maximum penalized likelihood estimation and smoothed EM algorithms for positive integral equations of the first kind *Numerical Functional Analysis and Optimization* 17 737-754

[8] Engl W H, Hanke M and Neubauer A 1996 *Regularization of Inverse Problems* (Dordrecht: Kluwer Academic Publishers)

[9] Holmes R B 1975 *Geometric Functional Analysis and Its Applications* (New York: Springer-Verlag)

[10] Iusem A N 1991 Convergence analysis for a multiplicatively relaxed EM algorithm *Mathematical Methods in the Applied Sciences* 14 573-593

[11] Iusem A N 1992 A short convergence proof of the EM algorithm for a specific Poisson model *REBRAPE* 6(1) 57-67

[12] Kondor A 1983 Method of convergent weights An iterative procedure for solving Fredholm's integral equations of the first kind *Nuclear Instruments and Methods in Physics Research* 216 177-181

[13] Lange K and Carson R 1984 EM reconstruction algorithms for emission and transmission tomography *J Computer Assisted Tomography* 8 306-316

[14] Latham G A and Anderssen R S 1994 On the stabilization inherent in the EMS algorithm *Inverse Problems* 10 161-183

[15] McLachlan G J and Krishnan T 1997 *The EM Algorithm and Extensions* (New York: John Wiley & Sons)

[16] Morozov V A 1966 On the solution of functional equations by the method of regularization *Soviet Math. Dokl.* 7 414-417

[17] Mülthei H N and Schorr B 1987 On an iterative method for a class of integral equations of the first kind *Math. Methods Appl. Sci.* 9(2) 137-168

[18] Mülthei H N and Schorr B 1989 On properties of the iterative maximum likelihood reconstruction method *Math. Methods Appl. Sci.* 11 331-342

[19] Mülthei H N 1993 Iterative continuous maximum-likelihood reconstruction methods *Math. Methods Appl. Sci.* 15 275-286

[20] Perlman M D 1974 Jensen's inequality for a convex vector-valued function on an infinite-dimensional space *J. Multivar. Anal.* 4 52-65

[21] Resmerita E and Anderssen R S 2007 A joint additive Kullback-Leibler residual minimization and regularization for linear inverse problems *Mathematical Methods in the Applied Sciences* 1527-1544

[22] Shepp L A and Vardi Y 1982 Maximum Likelihood Reconstruction in Positron Emission Tomography *IEEE Trans. Medical Imaging* 1(2) 113-122

[23] Silverman B, Jones M, Wilson J and Nychka D 1990 A Smoothed EM Approach to Indirect Estimation Problems, with Particular Reference to Stereology and Emission Tomography (with Discussion) *J. Royal Statistical Soc.* Series B (52) 271-324

[24] Vardi Y and Lee D 1993 From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems *J. R. Stat. Soc.* Ser. B Methodol. 55(3) 569-612

[25] Vardi Y, Shepp L A and Kaufmann L 1985 A statistical model for positron emission tomography *J. Am. Stat. Assoc.* 80 8-37

[26] Veklerov E and and Llacer J 1987 Stopping rule for the MLE algorithm based on statistical hypothesis testing *IEEE Trans. Med. Imag.* 6(4) 313-319

[27] Yosida K 1975 *Functional Analysis* (New York: Springer)