Majorizing-Functions Interior Point Methods for Solving Large Scale Regularized Maximum Likelihood Problems*

Alvaro R. De Pierro

State University of Campinas
Department of Applied Mathematics
CP 6065, CEP 13081-970, Campinas, SP, Brazil

Abstract

The Expectation Maximization (EM) algorithm has been a very popular method for maximizing likelihoods in many statistical estimation problems. Recently, a new interpretation of the EM algorithm applied to Poisson likelihood maximization was given in terms of majorizing-functions (minorizing for the concave case); i.e., each iteration is defined by majorizing the convex function to be minimized by another function which is tangent at the current iteration point. This interpretation allows a natural extension of the method to regularized problems. In this article we present a further extension of the majorizing-functions approach to general likelihood functions and regularization terms. We prove convergence results.

Keywords: Maximum Likelihood, Regularization, Interior Point Methods, Majorizing Functions.

1 Introduction

Maximum Likelihood (ML) computation via the Expectation Maximization (EM) algorithm has become a very popular method for solving a very large class of statistical optimization problems. In particular, since it was first suggested by Shepp and Vardi [16] in 1982 for emission tomography, it has originated a great deal of research in image reconstruction, an important area of application.

The EM algorithm for general estimation problems was proposed in 1977 [5] by Dempster, Laird and Rubin and the basic idea is very simple. Suppose the observed

*This work was partially supported by CNPq grant nº 301699/81 and FAPESP grant nº 94/0012-6.
Received december 1996, revised version accepted june 1997.

data in some experiment or sequence of experiments is a random vector y with density function $g(Y,x)$, where x is some vector of parameters to be estimated. In general, it may be difficult to maximize $g(Y,x)$ with respect to x, especially if x is a very large vector, as it is the case in Emission Computed Tomography (ECT). A possible solution to this problem is to perform the thought experiment of embedding the sample space for Y in a richer or larger sample space $\overline{Y}$ where optimization problems are easier to solve. The observed data is a realization from Y (incomplete data). The corresponding $\overline{y}$ in $\overline{Y}$ is not observed directly but through g. Especially, we assume that there is a mapping $\overline{y} \stackrel{h}{\rightarrow} y(\overline{y})$ from $\overline{Y}$ to Y, and that $\overline{y}$ is known only to lie in $\overline{Y}(y)$, the subset of $\overline{Y}$ determined by the equation $y = y(\overline{y})$, where y is the observed data, $\overline{y}$ will be refered as the complete data.

If $\overline{Y}$ has a density function $f(\overline{Y},x)$ with respect to some measure $\mu(\overline{Y})$, $g(Y,x)$ can be recovered by integration, i.e.,

$$g(Y,x) = \int f(\overline{Y},x)d\mu(\overline{Y}).\{\overline{Y} = h(\overline{Y}) = Y\} \tag{1}$$

Our problem is now finding a value of x which maximizes $g(Y,x)$ given an observed y; in other words:

$$\max_{x\in\Omega} L(x) = \log g(Y,x), \tag{2}$$

where Ω is a given convex set in $\mathbb{R}^n$.

If

$$k(\overline{Y}/Y,x) = \frac{f(\overline{Y},x)}{g(Y,x)} \tag{3}$$

is the conditional density of $\overline{Y}$ given Y and x, $L(x)$ can be written in the form

$$L(x) = \log f(\overline{Y},x) - \log k(\overline{Y}/Y,x). \tag{4}$$

We define

$$H(\widetilde{x},x) = E(\log k(\overline{Y}/Y,x)/Y,x) \tag{5}$$

where $E(\cdot/Y,x)$ denotes the expectation given Y and x. Then

$$Q(\widetilde{x},x) = L(\widetilde{x}) + H(\widetilde{x},x). \tag{6}$$

The following property of H motivates the EM algorithm and its generalizations.

Lemma. $\forall(\widetilde{x},x)$ in Ω^2

$$H(\widetilde{x},x) \leq H(x,x) \tag{7}$$

with equality iff $k(\overline{Y}/Y,\widetilde{x}) = k(\overline{Y}/Y,x)$ almost everywhere.

Proof. See [5], Lemma 1. □

(7) is a straightforward consequence of Jensen's inequality and the Lemma essentially says that $H(\widetilde{x}, x)$ has a maximum for $\widetilde{x} = x$ as a function of the first variable.

Using (6), for a given x^k and for all x we have that

$$L(x) - L(x^k) = [Q(x, x^k) - Q(x^k, x^k)] + [H(x^k, x^k) - H(x, x^k)]. \tag{8}$$

By the Lemma the second term between brackets is always nonnegative, then if we choose x such that

$$Q(x, x^k) \geq Q(x^k, x^k). \tag{9}$$

$L(x)$ will be greater or equal than $L(x^k)$. Therefore the function L will be nondecreasing and this is the first step to develop an algorithm for maximizing L. Taking into account this property we define the

EM Algorithm: Given $x^0 \in \Omega$, for $k = 0,\ 1,\ 2 \ldots$

E-step: Compute the conditional expectation

$$E(\log f(\overline{Y}, x)/Y, x^k) = Q(x, x^k); \tag{10}$$

M-step: Choose x^{k+1} to be

$$\arg\max_{x \in \Omega} Q(x, x^k). \tag{11}$$

In the particular case of Emission Computed Tomography (ECT), the ML approach mathematically consists of estimating the emission densities x_j (for each pixel [picture element] j) by solving the optimization problem

$$\max_{x \geq 0}\ L(x) = \sum_{i=1}^{m} y_i \log\langle a_i, x\rangle - \langle a_i, x\rangle, \tag{12}$$

where $y_i (i = 1, \ldots, m)$ represents the total emission counts for the i-th pair of detectors (if it is a PET [Posotron Emission Tomography] reconstruction [17]), $a_i = \{a_{ij}\} j = 1, \ldots, n)$ is the vector derived from the discretization of the i-th line integral, $x = (x_1, \ldots, x_n)^t$ is the image vector of densities and $\langle , \rangle$ denotes the standard inner product. The projection matrix A with rows a_i is assumed to have nonzero rows (every projection line intersects at least one pixel) and columns (every pixel is intersected by at least one projection line).

If $L(x)$ is given by (12) and Y is the sample space of the emission data y, $\overline{Y}$ may be chosen as the sample space of vectors $\overline{y} = \{y_{ij}\}$, where y_{ij} is the number of emissions from pixel j detected by pair i; it is clear that

$$y_i = \sum_{j=1}^{n} y_{ij}, \tag{13}$$

and the expected value of y_{ij} (unknown) given y_i and x^k is

$$E(y_{ij}/y_i, x^k) = \frac{y_i a_{ij} x_j^k}{\langle a_i, x^k \rangle}. \tag{14}$$

The extended log-likelihood for this problem is

$$\overline{L}(x) = \sum_{ij} y_{ij} \log a_{ij} x_j - a_{ij} x_j, \tag{15}$$

and applying the expectation given y and x^k to $\overline{L}$ we obtain

$$Q \equiv Q_1(x, x^k) = \sum_{i,j} \frac{y_i a_{ij} x_j^k}{\langle a_i, x^k \rangle} \log a_{ij} x_j - a_{ij} x_j, \tag{16}$$

using (14) and linearity. In the same way

$$H \equiv H_1(x, x^k) = \sum_{i,j} \frac{y_i a_{ij} x_j^k}{\langle a_i, x^k \rangle} \log a_{ij} x_j - \sum_{i=1}^{m} y_i \log \langle a_i, x \rangle \tag{17}$$

and the sequence defined by (11) is, for $k = 0,\ 1,\ 2, \ldots$.

$$x_j^{k+1} = x_j^k A_j^k, \tag{18}$$

where

$$A_j^k = \sum_{i=1}^{m} \frac{y_i a_{ij}}{\langle a_i, x^k \rangle}. \tag{19}$$

Analyzing the EM algorithm in this particular case in a non-statistical framework we can make the following observation: the expectation step substitutes the original problem of maximizing L by another simpler problem of maximizing Q whose variables are separated. This is done, using the fact that $\langle a_i, x \rangle$ can be represented as a convex combination of the x_j with coeffficients $\dfrac{a_{ij} x_j^k}{\langle a_i, x^k \rangle}$ and the concavity of $\log x$. Moreover, it is worth observing that, essentially, the EM algorithm consists of building at each step a simpler function that minorizes L and coincides with L and its gradient at the current iterate. The maximization of this simpler function generates the next point. This observation led us to generalize the EM algorithm for Poisson likelihoods to the case in which there is a regularization (penalization) term [7]. Numerical simulations for this algorithm can be found in [9]. In this article we present a further generalization to regularized ML problems that includes important examples that will be described later. The main statistical problem is to estimate parameters related by a nonnegative matrix A with data y, with a known distribution density, adding some prior information that is mainly contained in the regularization term.

So, our general optimization problem (we switch now to minimization of convex functions instead of maximization of concave ones, just because of this generalized habit in the optimization community) has the form

$$\min_{x \geq 0} G(x) = L(x) + R(x), \tag{20}$$

where

$$L(x) = \sum_{i=1}^{m} g_i(\langle a_i, x \rangle), \tag{21}$$

$$R(x) = \sum_{l=1}^{p} \hat{g}_l(\langle s_l, x \rangle), \tag{22}$$

$g_i (i = 1, \ldots, m)$ and $\hat{g}_l (l = 1, \ldots, p)$ are single variable real valued strictly convex functions, the $\hat{g}_l$'s continuously differentiable in the real line and the g_i's at least in the positive line. A is a nonnegative matrix as previously defined and the s_l are the rows of the real pxn matrix S with elements s_{lj}.

Also we will assume the following properties for the g's,

$$lim_{t \to 0^+} \; g_i'(t) \; < \; 0, \tag{23}$$

$$\hat{g_l}'(0) \; \leq \; 0, \tag{24}$$

and

$$lim_{t \to \infty} \; g_i(t) \; = \; +\infty. \tag{25}$$

Several important likelihood functions g_i and regularization functions $\hat{g}_l$ satisfy the conditions above. One important family of g_i's can be derived from the so called ϕ-divergences [2], defined in the following way. Let ϕ be a convex differentiable real valued function defined in R_+ and satifying some boundary conditions (see [2]); then for every x and y in R_+^m we define the divergence measure

$$d_\phi(x, y) = \sum_{i=1}^{m} y_i \phi(\frac{x_i}{y_i}). \tag{26}$$

Now, let $y = (y_i)$, and consider

$$g_i(t) = y_i \phi(\frac{t}{y_i}). \tag{27}$$

If $\phi(t) = -\ln t + t - 1$, we retrieve the Kullback-Leibler consistency measure, or equivalently, the Poisson likelihood, and the algorithm is the modified EM of [7]. $\phi(t) = (1 - \sqrt{t})^2$ gives the Hellinger 'distance'. If $\phi(t) = (t-1)^2$, we have the χ^2 divergence and the resulting algorithm is the modification of ISRA (Iterative Space Reconstruction Algorithm) for regularized nonnegative least squares problems presented in [6]. Other important examples of ϕ-divergences with their properties can

be found in [2]. Other examples, not being derived in [2] in [12].

Regarding the $\hat{g}_l$'s, a very important case was proposed in [8], and a list of other possibilities is given in [11].

In the next section we present an interior point algorithm for solving the general problem (20), in section 3 we prove some convergence properties of the new algorithm and, finally, in section 4 we present some concluding remarks.

2 The Majorizing-Functions Algorithm

Before introducing the new algorithm we need first to define the majorizing functions via convexity.

Using the convexity of the g_i's we obtain

$$g_i(\langle a_i, x\rangle) = g_i(\sum_{j=1}^{n} \frac{a_{ij}x'_j}{\langle a_i, x'\rangle x'_j} x_j \langle a_i, x'\rangle) \tag{28}$$

$$\leq \sum_{j=1}^{n} \frac{a_{ij}x'_j}{\langle a_l, x'\rangle} g_i(\frac{\langle a_i, x'\rangle x_j}{x'_j}). \tag{29}$$

And we define

$$\Phi^L(x, x') = \sum_{j=1}^{n} \Phi_j^L(x_j, x'), \tag{30}$$

where

$$\Phi_j^L(x_j, x') = \sum_{i=1}^{m} \frac{a_{ij}x'_j}{\langle s_l, x'\rangle} g_i(\frac{\langle s_l, x'\rangle x_j}{x'_j}). \tag{31}$$

Now let λ_j^l $(l = 1, \ldots, p;\ j = 1, \ldots, n)$ be nonnegative real numbers such that

$$\sum_{j=1}^{n} \lambda_j^l = 1, \quad \text{for} \quad l = 1, \ldots, p, \tag{32}$$

if

$$c_j^l = s_{lj}/\lambda_j^l, \quad \text{for} \quad j = 1, \ldots, n,\ l = 1, \ldots, p, \tag{33}$$

(If $s_{lj} = 0$, we define $\lambda_j^l = c_j^l = 0;\ \lambda_j^l \neq 0$ in any other case), and for a given x^k

$$d_j^{lk} = \begin{cases} \langle s_l, x^k\rangle - \dfrac{1}{\lambda_j^l} s_{lj} x_j^k & \text{if} \quad \lambda_j^l \neq 0 \quad \text{for} \quad j = 1, \ldots n \\ \\ \langle s_l, x^k\rangle & \text{otherwise} \end{cases} \tag{34}$$

Using (29) and the convexity of the $\hat{g}_l$'s, we get

$$\begin{aligned} \hat{g}_l(\langle s_l, x\rangle) &= \hat{g}_l(\sum_{j=1}^{n} \lambda_j^l (c_j^l x_j + d_j^{lk}) \\ &\leq \sum_{j=1}^{n} \lambda_j^l \hat{g}_l(c_j^l x_j + d_j^{lk}). \end{aligned} \tag{35}$$

$$\tag{36}$$

And we define

$$\Phi^R(x, x') = \sum_{j=1}^{n} \Phi_j^R(x_j, x'), \tag{37}$$

where

$$\Phi_j^R(x_j, x') = \sum_{l=1}^{p} \lambda_j^l \hat{g}_l(c_j^l x_j + d_j^{lk}). \tag{38}$$

Finally we define

$$\Phi(x, x') = \Phi^L(x, x') + \Phi^R(x, x'). \tag{39}$$

The following properties are easily verified and they motivate the definition of the algorithm.

Proposition 1. If R_+^n denotes the positive orthant, $\forall (x, x') \in (R_+^n)^2$: (i) $G(x) \leq \Phi(x, x')$, (ii) $G(x') = \Phi(x', x')$, (iii) $\nabla G(x') = \nabla \Phi(x', x')$.

Proof. (i) is an imediate consequence of the inequalities (29) and (35) above. (ii) and (iii) are straightforward calculations. ■

The Algorithm: Given $x^0 \geq 0$ and for $k = 0, 1 \ldots,$

$$\tilde{x}^k = \arg\min_{x \geq 0} \quad \Phi(x, x^k), \tag{40}$$

anf for a given $0 < \beta < 1$,

$$x_j^{k+1} = \begin{cases} \beta x_j^k & \text{if } \tilde{x}_j^k = 0 \\ \tilde{x}_j^k & \text{otherwise.} \end{cases} \tag{41}$$

From (40),it is clear that ,for $j = 1, \ldots, n$,

$$\tilde{x}_j^{k_s} \nabla \Phi(\tilde{x}^{k_s}, x^{k_s})_j = 0, \tag{42}$$

that

$$\nabla \Phi(\tilde{x}^{k_s}, x^{k_s})_j \geq 0 \tag{43}$$

and that $\tilde{x}^{k_s}$ is ≥ 0. On the other hand x^{k+1} is strictly positive $\forall k$, so, the algorithm is well defined.

3 Convergence Results

In this section we will prove convergence of algorithm (40-41) assuming uniqueness of the solution, a very natural assumption when dealing with regularized problems. Also we will assume that $\forall\ x$ and x' $x_j\nabla\Phi(x_j,x')_j$ is a continuous function of x_j in the nonnegative orthant.

Lemma 2. For $k=1,2,\ldots$

$$G(x^{k+1})=\Phi(x^{k+1},x^{k+1})\leq\Phi(x^{k+1},x^k)\leq\Phi(x^k,x^k)=G(x^k). \tag{44}$$

Proof. The equalities follow from Proposition 1(ii), the first inequality is a consequence of Proposition 1(i) and (ii), and the second one of the definition of the algorithm. ■

Corollary 3. The sequences $\{G(x^k)\}$ and $\{\Phi(x^{k+1},x^k)\}$ converge to the same limit G^*.

Proof. This is consequence of the fact that $G(x)$ is bounded above (remember that the problem has a solution) and (44).

Corollary 4. $\{x^k\}$ is bounded.

Proof. G is strictly convex and because of (25) it has bounded level sets, so, Lemma 2 implies that $\{x^k\}$ is bounded. ■

Proposition 5. Every limit point $\widehat{x}$ of (40-41) satisfies

$$\widehat{x}_j\nabla G(\widehat{x})_j=0,\quad j=1,\ldots,n. \tag{45}$$

Proof. Let $\{x^{k_s}\}$ be a subsequence of (40-41) convergent to $\widehat{x}$ and suppose that x^{k_s+1} converges to $\bar{x}$ (it should be a subsequence of x^{k_s+1} but we will denote it as being the same to simplify the notation), and that $\tilde{x}^{k_s}$ converges to $\tilde{x}$.

Suppose now that, for some $u\in[1,n]$

$$\widehat{x}_u\nabla G(\widehat{x})_u\neq 0. \tag{46}$$

Because of Lemma 2 and the continuity condition $\widehat{\widehat{x}}=Alg(\widehat{x})$ is well defined, where $Alg(x)$ denotes the application of the algorithm (40-41) to the point x. Using the continuity in equation (42), we take limits and get

$$\tilde{x_j}\nabla\Phi_j(\tilde{x_j},\widehat{x})=0,\quad \text{for}\quad j=1,\ldots,n. \tag{47}$$

If $\tilde{x_j}>0$, then $\nabla\Phi_j(\tilde{x_j},\widehat{x})=0$, and

$$\Phi_j(\tilde{x_j},\widehat{x})\leq\Phi_j(\widehat{\widehat{x}}_j,\widehat{x}); \tag{48}$$

but $\Phi_j(.,\widehat{x})$ is strictly convex and it has a unique minimum, so $\tilde{x}_j = \widehat{\widehat{x}}_j$. If $\tilde{x}_j = 0$ and $\widehat{x}_j$ is positive, for k_s large enough $\nabla\Phi_j(\tilde{x}_{k_s}, \widehat{x}_{k_s}) < 0$, using (23) and (24); but this contradicts the definition of $\tilde{x}_{k_s}$, therefore $\widehat{x}_j$ should be zero. It is straightforward deducing that $\bar{x} = \widehat{x}$ because, by definition, $\tilde{x}_j^{k_s} \leq x_j^{k_s+1} \leq \widehat{x}_j^{k_s}$. So, if $\widehat{x}_j \neq 0$, then $\tilde{x}_j \neq 0$, in particular

$$\Phi_u(\tilde{x}_u, \widehat{x}) < \Phi_u(\widehat{x}_u, \widehat{x}). \tag{49}$$

Also, by definition $\Phi_j(x_j^{k_s+1}, x^{k_s}) \geq \Phi_j(x_j^{k_s}, x^{k_s})$ for every j, and taking limits

$$\Phi_j(\tilde{x}_j, \widehat{x}) \leq \Phi_j(\widehat{x}_j, \widehat{x}). \tag{50}$$

From (49) and (50) and the definition of Φ it follows that

$$G(\tilde{x}) < G(\widehat{x}), \tag{51}$$

a contradiction with Corollary 3. ■

Theorem 6. With the hypotheses of the beginning of this section the sequence generated by (40-41) converges to a solution of (20).

Proof. By Proposition 5, if $\widehat{x}$ is a limit point of (40-41)

$$\widehat{x}_j \nabla G(\widehat{x})_j = 0 \quad \text{and} \quad \widehat{x}_j \geq 0. \tag{52}$$

So, we need only to prove that

$$\nabla G(\widehat{x})_j \geq 0, \tag{53}$$

for $\widehat{x}$ to satisfy the Kuhn-Tucker optimality conditions [1]. Using the same arguments of Theorem 1 in [7] or [10], we can prove that the whole sequence converges to $\widehat{x}$; i.e.; that there is only on limit point. Consider two such points x^* and x^{**}, and the sets (see [10])

$$N = \{1, 2, \ldots, n\}, \tag{54}$$

$$Z^* = \{j \in N : x_j^* = 0\}, \tag{55}$$

$$Z^{**} = \{j \in N : x_j^* = 0\}. \tag{56}$$

Let $G_{S^*}(x)$ be the restriction of $G(x)$ to the set

$$S^* = \{x : x_j = 0 \text{ for } j \in Z^*\}.$$

$G_{S^*}(x)$ is strictly convex in S^* and has a unique stationary point which is necessarily the minimum. Therefore, if $Z^* = Z^*, x^*$ and x^{**} should be the same. The number of limit points is bounded by the number of subsets of N, that it is finite. Now we can use Ostrowski's theorem [14] (the set of limit points of a sequence $\{x^*\}$ such that $x^{k+1} - x^k \underset{k\to\infty}{\longrightarrow} 0$ is connected) and deduce that $x^k \underset{k\to\infty}{\longrightarrow} x^*$. If $\widehat{x}_j > 0$, the result is obvious. Now assume that for some j, $\widehat{x}_j = 0$ and $\nabla G(\widehat{x})_j < 0$. In that case there exists a natural number K such that for $k \geq K$ it holds that $\nabla G(x^k)_j = \nabla\Phi(x_j^k, x^k)_j < 0$, but this implies that $x_j^{k+1} > x_j^k$ for $k \geq K$. A contradiction. ■

4 Concluding Remarks

In this article we have presented an extension of the EM algorithm to a very general class of problems arising from nonnegative likelihood maximization with concave 'a priori' information. Theorem 6 also extends and improves on some previously existing results for problems arising in image reconstruction ([6], [7]) as well as in image processing ([3], [12]). Further research is needed mainly in two directions: how to generate a more general family of majorizing functions (for example, the case in which convexity is used to separate the variables in blocks of more than one variable) and how to accelerate the resulting algorithms that can be sometimes rather slow (for example, by using some kind of decomposition like in [4]).

References

[1] M. Avriel, *Nonlinear Programming: Analysis and Methods*. Englewood Cliffs, NJ: Prentice Hall, 1976.

[2] A. Ben-tal, A. Ben-israel and M. Teboulle, "Certainty equivalents and information measures: duality and extremal principles". *J. Math. Analysis Appls.*, vol.157, pp. 211-236, 1991.

[3] C.A. Bouman and K. Sauer, "A unified approach to satistical tomography using coordinate descent optimization". *IEEE. Trans. Image Proc.*, vol. 5, nº 3, pp. 480-492, 1996.

[4] J. Browne and A.R. De Pierro, " A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography". *IEEE Trans. Med. Imaging*, vol.15, nº 5, pp. 687-700, 1996.

[5] A.P. Dempster, N.M. Laird and D.D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm". *J. Roy. Stat. Soc.*, Series B, vol. 39, pp. 1-38, 1977.

[6] A.R. De Pierro, "On the relation between ISRA and the EM algorithm for positron emission tomography". *IEEE Trans. Med. Imaging*, vol. 12, nº 2, pp. 328-333, 1993.

[7] A.R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography". *IEEE Trans. Med. Imaging*, vol. 14, nº 1, 1995.

[8] P. Green, "Bayesian reconstruction for emission tomography data using a modified EM algorithm". *IEEE Trans. Med. Imaging*, vol. 9, nº 1, pp. 84-93, 1990.

[9] G.T. Herman, A.R. De Pierro and N. Gai, "On methods for maximum a posteriori image reconstruction with a normal prior". *J. Visual Comm. Image Rep.*, vol. 3, nº 4, pp. 316-324, 1992.

[10] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography". *J. Comput. Assisted Tomog.*, vol. 8, pp. 306-316, 1987.

[11] K. Lange, M. Bahn and R. Little, "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography". *IEEE Trans. Med. Imaging*, vol.6, nº 1, 106-114, 1987.

[12] K. Lange and J.A. Fessler, "Globally convergent algorithms for maximum a posteriori transmission tomography". *IEEE Trans. Image Proc.*, vol. 4, nº 10, pp. 1430-1438, 1995.

[13] E. Levitan and G.T. Herman, "A maximum a posteriori probability expectation maximization algorithm for image reconstruction in emission tomography". *IEEE Trans. Med. Imaging*, vol. 6, nº 3, pp. 185-192, 1987.

[14] A.M. Ostrowski, *Solution of Equations in Euclidean and Banach Spaces*. New York, Academic Press, 1973.

[15] A. Rockmore and A. Macovski, "A maximum likelihood approach to emission image reconstruction from projections". *IEEE Trans. Nucl. Sci.*, vol. 23, pp. 1428-1432, 1976.

[16] L. Sheep and Y. Vardi, "Maximum likelihood reconstruction for emission tomography". *IEEE Trans. Med. Imaging*, vol. pp. 113-121, 1982.

[17] M.M. Ter-pogossian, M. Raichle and B.E. Sobel, "Positron Emission Tomography". *Scientif. Amer.*, vol. 243, nº 4, pp. 170-181, 1980.