Likelihood Maximization for List-Mode Emission Tomographic Image Reconstruction

Charles Byrne, Member, IEEE

Abstract—The maximum a posteriori (MAP) Bayesian iterative algorithm using priors that are gamma distributed, due to Lange, Bahn and Little, is extended to include parameter choices that fall outside the gamma distribution model. Special cases of the resulting iterative method include the expectation maximization maximum likelihood (EMML) method based on the Poisson model in emission tomography, as well as algorithms obtained by Parra and Barrett and by Huesman et al. that converge to maximum likelihood and maximum conditional likelihood estimates of radionuclide intensities for list-mode emission tomography. The approach taken here is optimization-theoretic and does not rely on the usual expectation maximization (EM) formalism. Block-iterative variants of the algorithms are presented. A self-contained, elementary proof of convergence of the algorithm is included.

Index Terms—Emission tomography, image reconstruction, iterative algorithm, probabilistic mixture.

I. INTRODUCTION

In bin-mode single-photon emission computed tomography (SPECT) photons are detected at one of \( I \) predetermined detector locations. During the course of the scanning, most detector locations count multiple arrivals, the counts at each detector are modeled as independent Poisson random variables, and likelihood maximization is used to estimate the radionuclide intensity levels at each voxel [29], [19].

For positron emission tomography (PET) coincidence imaging, a detection corresponds to the (nearly) simultaneous arrival of two photons at opposite ends of a line segment, or line of response, joining a pair of detector locations. The set of all such pairs is quite large, and in the course of the scanning, most of these pairs are associated with no detected events, some with one and typically none with more than one. The bin-mode approach can introduce distortions in the imaging by lumping together detected events that differ significantly, thereby ignoring potentially useful information. Because it is highly inefficient to store a large number of zero values and a smaller number of values that are one, it is common to use list-mode processing, whereby one maintains a list of those pairs associated with a detected event, possibly including additional information, such as the energies of the detected pairs of photons. Once again, likelihood maximization is used to estimate the radionuclide intensities at the voxels. Now, however, the Poisson model is inappropriate to describe the list and the likelihood function is that of a probabilistic mixture [27].

In both of these situations, the expectation maximization (EM) method [13] can be used to derive an iterative procedure to maximize the likelihood, as was done for bin-mode by Shepp and Vardi [29] and Lange and Carson [19], and for list-mode by Parra and Barrett and by Huesman et al. [16]. In these papers, the EM formalism was employed to obtain an iterative method for likelihood maximization.

In this paper, we avoid the EM approach by rewriting the likelihood function as a penalized cross-entropy and applying our algorithm for minimizing such functions. Our algorithm includes, as particular cases, the EM algorithm for SPECT of Shepp and Vardi [29] and Lange and Carson [19], called here the EMML algorithm, the Bayesian maximum a posteriori (MAP) method of Lange et al. [20], and the likelihood maximization method of Parra and Barrett [26] and of Huesman et al. [16]. In fact, our iterative algorithm is an extension of the MAP method for regularized likelihood maximization given in [20], involving gamma-distributed prior distributions on the voxel intensities. While formally identical to their MAP method, ours allows the parameters to take on nonpositive values, which are outside the gamma distribution model. With this added flexibility, we obtain the Parra and Barrett algorithm as a particular case. Negative parameter values can arise when we use probability density functions (pdf's) to describe the distribution of possible events, rather than probability functions (pf’s).

The Poisson model is used only to describe the total counts, as in [26]. The bin-mode EM algorithm for SPECT can be derived in this way, without the additional assumptions that the individual counts at each detector are Poisson distributed. Our approach is entirely optimization-theoretic and does not involve the EM formalism.

In Section II, we give a short history of recent efforts to prove convergence of the EM algorithm and its various relatives, both to motivate the inclusion of yet another proof here, and to put our proof in some context. List-mode emission tomography and likelihood maximization are treated in Section III. In Section IV, we present a brief overview of the iterative algorithms we shall be discussing. By viewing them together we can better see the similarities and subtle differences that will be important later in the discussion. Section V concerns our new iterative algorithm, with particular cases of the algorithm discussed in Section VI. In Section VII, we look at the relationship between the iterative EM algorithm for list-mode processing and the EMML.
bin-mode algorithm. A block-iterative version of our algorithm
is presented in Section VIII. Underlying the formulation of
the likelihood function for the list-mode case is the notion of a proba-
blistic mixture; in Section IX we give a simple example of
such a mixture and show how our algorithm can be applied to
estimate the mixing coefficients. A self-contained and elemen-
tary proof of convergence for our algorithm is given in the Ap-
pendix.

II. A BRIEF HISTORY OF CONVERGENCE PROOFS FOR THE
EM ALGORITHM

It seems customary to credit the EM algorithm to Dempster,
Laird and Rubin [13], although it is made clear in that paper
that the EM algorithm did not originate with them. For detailed
histories of the EM algorithm, see the paper by Meng and van
dyk [24] and the recent text by McLachlan and Krishnan [23].

The EM algorithm concerns a parameterized family of pdf's
or pf's \( f\left(x|x^k\right) \), where \( x \) is a parameter vector. The algorithm
generates a sequence \( \{x^k\} \) that we hope will converge to a pa-
rameter vector maximizing the likelihood. The tricky point is
to relate convergence of the sequence of functions \( f\left(x|x^k\right) \)
to convergence of the sequence \( \{x^k\} \). The convergence proof
in [13] contained an error involving just this point. Wu [31]
pointed out this error and provided sufficient conditions for con-
vergence, in terms of a forcing function relating the behavior of
the \( f\left(x|x^k\right) \) to that of the \( x \) (see also Redner and Walker [27]).

In [29], Shepp and Vardi consider the special case of the EM
algorithm applied to likelihood maximization in emission tomography; we
call this special case the EMML algorithm in this paper. As
Lange and Carson note in [19], the convergence proof in [29]
has a gap, in that it ignores the possibility that the limit vector
of the iterative sequence may have zero entries. The proof given
by Lange and Carson is valid, but requires the additional
assumption that the likelihood maximizer be unique. In [30],
the authors acknowledge the point made by Lange and Carson and
correct the earlier proof using results of Csiszar and Tusnady

There is a misstatement in the theorem in [30], however. They
assert (incorrectly) that the likelihood has a unique maximizer if
and only if the rows of the matrix of probabilities involved span
\( R^J \), where \( J \) is the number of voxels; for the bin-mode case this
implies that there are at least as many bins as there are voxels or
pixels. If they were correct, the typically noisy maximum like-
lihood images could be smoothed by using more voxels than
bins, to guarantee nonunique likelihood maximizers, generating
a number of distinct noisy maximum likelihood images from
different starting points, and then averaging them, to obtain a
smoother maximum likelihood image. However, as shown in
[4], smooth maximum likelihood images cannot be obtained in
this way, since the maximum likelihood estimate is almost al-
tways unique for noisy data, hence independent of starting point,
even when there are more voxels than bins (see also [5]).

To obtain smoother images, regularization is often used. An
example is the maximum \( a \ posteriori \) Bayesian approach of
Lange et al. [20], involving prior gamma distributions for the
unknowns. Since the mean and variance of the gamma distribu-
tion are positive, they do not have any problem with limit vectors
having any zero entries.

As noted in [30], the article [12] of Csiszar and Tusnady can
difficult to follow. A short proof of convergence for the
EMML algorithm was given by Iusem in [17], using the joint
convexity of the KL distance. A more general result along the
same lines is given in [9]. A shorter and more elementary proof
of convergence of the EMML was given in [6]. The convergence
proof given in this paper is an extension of the proof for EMML
in [6].

III. OVERVIEW OF THE RELEVANT ALGORITHMS

Central to our discussion is the notion of cross-ent-
ropy or Kullback–Leibler (KL) distance [18] between vectors
with nonnegative entries. For \( a > 0 \) and \( b > 0 \), let
\( KL(a, b) = a \log(a/b) + b - a \). For \( KL(0, b) = b \) be defined component-wise as
\( KL(a, b) = a \log(a/b) + b - a \). The EMML algorithm for SPECT [29], [19] can be viewed as an
iterative procedure for finding a nonnegative (approximate)
solution \( x \) for a system of linear equations of the form
\( y = Px \), where \( y = (y_1, \ldots, y_N)^T \) is a vector with positive entries and
\( P \) is a matrix with entries \( P_{nj} > 0 \), for \( n = 1, \ldots, N \) and
\( j = 1, \ldots, J \), such that \( s_j = \sum_{n=1}^N P_{nj} > 0 \) for each \( j \). The
EMML algorithm converges to a minimizer of the function
\( KL(y, Px) \) over all nonnegative vectors. With \( (Px)_n = \frac{\sum_{j=1}^J P_{nj} x_j}{s_j} \), the iterative step of the EMML is

\[
\begin{align*}
x_j^{k+1} &= s_j^{-1} x_j^k \sum_{n=1}^N P_{nj} \frac{y_n}{(Px_n)^n}.
\end{align*}
\]

When the EMML algorithm is used to reconstruct images from
noisy data, the result is often too rough to be useful. Variations of
the EMML involving regularization are then employed to obtain
smoother images.

One such regularization method is the maximum \( a \ posteriori \)
(MAP) Bayesian method of Lange et al. [20], in which each
unknown entry \( x_{ij} \) is assumed to have a prior gamma distribu-
tion, with mean \( \mu_j > 0 \) and variance \( \sigma_j^2 \) assumed to lie within the interval \((0, \mu_j)\). The function to be minimized can be written as

\[
G(x) = KL(y, Px) + \sum_{j=1}^J \beta_j KL(\gamma_j, x_j)
\]

where \( \beta_j = \mu_j / \sigma_j^2 \) and \( \gamma_j = (\mu_j - \sigma_j^2) / \mu_j > 0 \). The
iterative step of their MAP method is

\[
\begin{align*}
x_j^{k+1} &= \frac{\alpha_j s_j^{-1} x_j^k \sum_{n=1}^N P_{nj} \frac{y_n}{(Px_n)^n} + (1 - \alpha_j) \gamma_j}{s_j^{-1} x_j^k \sum_{n=1}^N P_{nj} \frac{y_n}{(Px_n)^n}}.
\end{align*}
\]
with \( \alpha_j = s_j/(s_j + \beta_j) \). Note that if we violate the restrictions on the parameters \( \gamma_j \) and take \( \gamma_j = 0 \) for each \( j \), then the function to be minimized becomes

\[
G(x) = KL(y, Px) + \sum_{j=1}^{J} \beta_j x_j. \tag{3.4}
\]

As we shall see, maximizing the likelihood in list-mode image processing in SPECT and PET is equivalent to minimizing a function of the form in (3.4).

In the list-mode case, the vector \( y \) is the vector \( u \) whose entries are all one and the \( \beta_j \) may sometimes be negative, which is not permitted in the gamma distribution. The iterative algorithm for this case has the following update equation

**The List-Mode EM Algorithm:**

\[
x_j^{k+1} = d_j^{-1} x_j^k \sum_{n=1}^{N} P_{nj} \frac{1}{(P x)^n} \tag{3.5}
\]

with \( d_j \in (0, 1] \) denoting the probability that an emission at the \( j \)th voxel will be detected, and \( \beta_j = d_j - s_j \). This is the algorithm for list-mode processing given by Parra and Barrett [26] and by Huesman et al. [16]. When pdf’s are used to describe the probability distribution of events, the entries of the matrix \( P \) are values of the pdf, so the \( s_j \) can be any positive numbers, while the \( d_j \) are restricted to \([0, 1]\).

The main goal in this paper is to derive and prove convergence of an extension of the MAP method in (3.3) that is sufficiently general to include the list-mode EM algorithm. Specifically, we shall prove that the iterative algorithm with the update equation.

**Our Algorithm:**

\[
x_j^{k+1} = \alpha_j s_j^{-1} x_j^k \sum_{n=1}^{N} P_{nj} \frac{y_n}{(P x)^n} + (1 - \alpha_j) \gamma_j \tag{3.6}
\]

with \( \alpha_j = s_j/(s_j + \beta_j) \), converges to a minimizer of the function

\[
G(x) = KL(y, Px) + \sum_{j=1}^{J} \beta_j KL(\gamma_j, x_j), \tag{3.7}
\]

provided \( s_j + \beta_j > 0 \) and \( \beta_j \gamma_j \geq 0 \) for all \( j \).

**IV. LIST-MODE PROCESSING**

In the list-mode approach, the set of possible outcomes is usually either infinite or so large that typically no single outcome occurs more than once and most never occur at all. Of course, we can, if we wish, consider bin-mode data in list-mode format; in that case, the set of possible outcomes is much smaller and most of them occur many times.

For \( j = 1, 2, \ldots, J \) let \( x_j \geq 0 \), be the expected number of events originating within the \( j \)th voxel during the scanning time; we assume that \( x_j \) is proportional to the intensity of the radionuclide within the \( j \)th voxel. Let \( f_j(\cdot) \) the pdf or pf describing the distribution of events originating in the \( j \)th voxel. With \( x_j = \sum_{j=1}^{J} x_j \), the probability that an event originated within the \( j \)th voxel, given that it originated within one of the \( J \) voxels, is \( x_j/x_+ \). Therefore, by the familiar rule of total probability, the distribution of random events is described by

\[
f(\cdot) = \frac{1}{x_+} \sum_{j=1}^{J} x_j f_j(\cdot). \tag{4.1}
\]

This indicates that the distribution of random events can be considered as a probabilistic mixture [27], with the mixing coefficients \( x_j/x_+ \) determined by the relative intensities of the radionuclides at each voxel. In a later section, we consider such mixtures in more detail.

We denote by \( \{z_1, \ldots, z_N\} \) the list of detected events. We assume that the items on the list were independently generated. The number \( N \) of events in the list is also a random quantity; to determine its distribution we need to discuss detectability, that is, the probability that an event originating within one of the voxels will be detected.

Let the set of possible events be denoted by \( Z \). If \( Z \) is an infinite set or a finite set so large that typically no single outcome occurs more than once and most never occur at all, the functions \( f_j(\cdot) \) are pdf or pf defined over \( Z \). We then let \( g(z) \in [0, 1] \) be the probability that \( z \in Z \) will be detected if it occurs. The probability of detecting an event originating within the \( j \)th voxel is assumed to be positive and given by \( d_j = \int f_j(z) g(z) dz \) or \( d_j = \sum_j f_j(z) g(z) \), depending on whether \( f_j(\cdot) \) is a pdf or a pf. Note that \( d_j \) does not depend on which events happen to be on the list.

For each \( j \), let \( d_j \) in \([0, 1]\) be the probability of detecting an event originating within the \( j \)th voxel. Then \( d \), the probability of detecting an event, is given by

\[
d = \frac{1}{x_+} \sum_{j=1}^{J} x_j d_j. \tag{4.2}
\]

We model the random number of items in the list as a Poisson random variable with mean \( x \cdot d \). Therefore, the probability of having \( N \) items in the list is

\[
p(N) = \exp(-x \cdot d)(x \cdot d)^N/N!. \tag{4.3}
\]

With \( x = (x_1, x_2, \ldots, x_J)^T \), the likelihood function \( L(x) \) is

\[
L(x) = p(N) \prod_{n=1}^{N} f(z_n). \tag{4.4}
\]

Taking logs of both sides and using (4.1), we get the log likelihood function, \( LL(x) \), given, except for constants, by

\[
LL(x) = -x \cdot d + N \log(x \cdot d) + \sum_{n=1}^{N} \log(P x)_n \tag{4.5}
\]

where \( P \) is the matrix with entries \( P_{nj} = f_j(z_n) \) and \((P x)_n = \sum_{j=1}^{J} P_{nj} x_j \). Note that

\[
\sum_{n=1}^{N} (P x)_n = \sum_{n=1}^{N} \sum_{j=1}^{J} P_{nj} x_j = \sum_{j=1}^{J} (\sum_{n=1}^{N} P_{nj}) x_j = \sum_{j=1}^{J} s_j x_j.
\]

To estimate the intensity vector \( x \) we maximize the function \( LL(x) \), subject to \( x_j \geq 0 \) for each \( j \). To obtain an iterative method for doing this, we rewrite the problem using the KL distance between nonnegative vectors.
With $u$ the vector whose entries are all one, we have

$$-LL(x) = KL(u, Px) - \sum_{n=1}^{N} (Px)_n + x_d d + \text{constants.}$$  \hfill (4.6)

With $s_j = \sum_{n=1}^{N} P_{nj} = \sum_{n=1}^{N} f_j(z_n)$, our problem is to minimize the function

$$F(x) = KL(u, Px) + \sum_{j=1}^{J} (d_j - s_j) x_j.$$  \hfill (4.7)

Note that since the functions $f_j(\cdot)$ may be pdf, $s_j$ can be any positive number, so $d_j - s_j$ need not be positive.

The function $F(x)$ that we wish to minimize, over all $x$ with $x_j \geq 0$, is a particular case of the function $G(x)$ given by

$$G(x) = KL(y, Px) + \sum_{j=1}^{J} \beta_j KL(\gamma_j, x_j)$$

where $y = (y_1, \ldots, y_N)^T$ is a vector with positive entries and $\gamma_j \geq 0$ for each $j$. To obtain $F(x)$ from $G(x)$ we set $\beta_j = d_j - s_j$ and $\gamma_j = 0$ for each $j$ and $y_n = u_n = 1$ for each $n$. In the Section V we derive an iterative algorithm to minimize $G(x)$ over all $x$ with nonnegative entries. The maximum likelihood algorithm for list-mode processing will then emerge as a particular case.

V. THE ITERATIVE ALGORITHM

Throughout this section we shall assume that $y$ is an $N$-dimensional vector with positive entries $y_n, n = 1, \ldots, N$ and let $y_+ = \sum_{n=1}^{N} y_n$. We also assume that $P$ is an $N$ by $J$ matrix with nonnegative entries, such that each row and column of $P$ has at least one positive entry. For $j = 1, 2, \ldots, J$, let $s_j = \sum_{n=1}^{N} P_{nj} > 0$ and choose $\gamma_j \geq 0$. For nonnegative vectors $x$ in the set $X = \{x \mid (Px)_n > 0, n = 1, \ldots, N\}$, let $G(x)$ be the function defined by (4.8). We consider now an iterative algorithm that minimizes $G$ over its domain, whenever $\beta_j + s_j > 0$ and $\beta_j \gamma_j \geq 0$ for each $j$. Using these conditions, we can rewrite the minimization problem as follows.

Let $d_j = \beta_j + s_j > 0$ and $c > 1$. Let $\beta_j$ be the matrix with entries $P_{nj} = c^{-1} P_{nj}$. Let $\beta_j = d_j - c^{-1} s_j$ and $\gamma_j$ such that $\beta_j \gamma_j \geq 0$. Then it is easily shown that

$$G(x) = KL(y, \hat{P} x) + \sum_{j=1}^{J} \beta_j KL(\gamma_j, x_j) + \text{constants.}$$  \hfill (5.1)

Consequently, by choosing $c$ large enough, we can make $\beta_j > 0$ and $\gamma_j \geq 0$ for each $j$. It is clear then that $G(x)$ attains its minimum value on its domain. This minimization problem is slightly more general than the one arising in the MAP method of [20], in that $\gamma_j = 0$ is permitted here. Our iterative algorithm is the following, which is formally the same as in [20]: with $x^0 > 0$ in $X$, $\hat{s}_j = \sum_{n=1}^{N} P_{nj}$ and $k = 0, 1, \ldots$

$$x_j^{k+1} = \hat{\alpha}_j \hat{s}_j^{-1} \sum_{n=1}^{N} \hat{P}_{nj} \frac{y_n}{(Px^k)_n} + (1 - \hat{\alpha}_j) \hat{\gamma}_j$$  \hfill (5.2)

with $\hat{\alpha}_j = \hat{s}_j/d_j$. In the notation of the original problem, the iterative step becomes

$$x_j^{k+1} = \alpha_j s_j^{-1} \sum_{n=1}^{N} P_{nj} \frac{y_n}{(Px^k)_n} + (1 - \alpha_j) \gamma_j$$  \hfill (5.3)

with $\alpha_j = s_j/d_j$. From (5.2), it is clear that $x_j^k > 0$ for each $j$ and each $k$. We can also see this from (5.3), since $\beta_j + s_j > 0$ and $\beta_j \gamma_j \geq 0$ for each $j$.

In the Appendix, we prove convergence of the iterative scheme in (5.3). We then have the following result:

**Theorem 5.1.** Let the following conditions hold: 1) $\beta_j \gamma_j > 0$ for each $j$ and 2) $s_j + \beta_j > 0$ for each $j$. Then the sequence generated by (5.3) converges to a nonnegative minimizer of the function $G(x)$ given by (4.8).

This theorem was proved in [20], using ideas from [19], for the case in which $N \geq 1, \beta_j > 0$ and $\gamma_j > 0$ for each $j$; a simplified proof was given in [4] for the special case of $\beta_j = \gamma_j = 0$ for all $j$. It is important to note that the theorem above permits $\gamma_j = 0$, and so this is slightly more general than the result in [20]; we shall need this when we derive the method of Parra and Barrett as a particular case of our algorithm.

Multiplying by $d_j = s_j + \beta_j$ on both sides of (5.3), we find that

$$\sum_{j=1}^{J} x_j^{k+1} = y_+ + \sum_{j=1}^{J} \beta_j \gamma_j.$$  \hfill (5.4)

We consider now several particular cases of this algorithm.

VI. SOME PARTICULAR CASES OF THE ALGORITHM

Several well known iterative reconstruction methods are particular cases of the algorithm given in the previous section.

A. The EMML Algorithm for SPECT

The general expectation maximization (EM) method for likelihood maximization was presented by Dempster, Laird and Rubin in [13]. Application of this general formalism to the specific case of emission tomography and Poisson data resulted in what, in medical imaging, is usually called the EMML (or, sometimes, the MLEM or just EM) algorithm [29], [30], [19], [20].

For $j = 1, \ldots, J$ and $i = 1, \ldots, I$, let $P_{ij} \geq 0$ be the probability that a photon emitted from the $j$th voxel is detected at the $i$th detector and let $Y_{ij}$ be the random number of such detections. The $Y_{ij}$ are assumed to be independent and Poisson distributed, with mean values $P_{ij} x_j$. The random number of detections made at the $i$th detector, denoted $Y_i$, is then Poisson distributed, with mean value $P x_i = (Px)_i = \sum_{j=1}^{J} P_{ij} x_j$. The probability that a photon emitted at the $j$th voxel will be detected is $d_j = s_j = \sum_{i=1}^{I} P_{ij}$. With $y_i > 0$ the actual
number of detections made at the \(i\)th detector, the likelihood function of the vector \(x\) becomes
\[
L(x) = \prod_{i=1}^{I} \exp(-P x_i) (P x_i)^{y_i} / y_i!
\]
(6.1)
and so maximizing the likelihood, subject to \(x_{j} \geq 0\), is equivalent to minimizing the function \(KL(y, P x)\), subject also to \(x_{j} \geq 0\). Applying our iterative algorithm, with \(x_{j}^{0} > 0\) and \(\beta_{j} = 0\) for each \(j\), we have
\[
x_{j}^{k+1} = s_{j}^{-1} x_{j}^{k} \sum_{i=1}^{I} P_{ij} \frac{y_{i}}{(P x_{j}^{k})_{i}}
\]
(6.2)
this is the EMML algorithm.

It is important to note that although this iterative method was originally derived as a special case of the Dempster, Laird and Rubin algorithm given in [13], it can be viewed apart from any statistical consideration, as an iterative method for finding a nonnegative vector \(x\) that minimizes \(KL(y, P x)\), for any positive vector \(y\) and any matrix \(P\) having nonnegative entries. Accordingly, whenever we encounter an iteration of the form (6.2) we can determine immediately that it is minimizing \(KL(y, P x)\).

In a later section, we shall rederive this algorithm, using only the assumption that the total counts are Poisson distributed, by converting list-mode data to bin-mode.

B. The MAP Method With Prior Gamma Distributions

Lange et al. [20] regularize the likelihood maximization problem in emission tomography using a Bayesian approach, in which each of the unknown values \(x_{j}\) is given a prior gamma distribution, with mean \(\mu_{j}\) and variance \(\sigma^{2}_{j}\). With \(\beta_{j} = \mu_{j} / \sigma^{2}_{j}\) and \(0 < \gamma_{j} = (\beta_{j} - \sigma^{2}_{j}) / \mu_{j}\), the likelihood function becomes
\[
G(x) = \frac{s_{j}^{-1} x_{j}^{k}}{\prod_{i=1}^{I} \left(\frac{y_{i}}{(P x_{j})_{i}}\right)}
\]
and the resulting method has the iterative step given in (5.3), with the restriction that \(\beta_{j} > 0\) and \(\gamma_{j} > 0\) for all \(j\).

C. The List-Mode EM Algorithm

As we saw earlier, the list-mode likelihood maximization problem is equivalent to minimizing the function
\[
F(x) = KL(u, P x) + \sum_{i=1}^{I} (d_{j} - s_{j}) x_{j}
\]
(6.3)
over vectors \(x\) with nonnegative entries. Since the functions \(f_{j}(\cdot)\) may be pdf, \(s_{j}\) can be any positive number, so \(\beta_{j} = d_{j} - s_{j}\) need not be positive; however, we do have \(\beta_{j} + s_{j} = d_{j} > 0\).

Applying our iterative method, we obtain the following:
\[
x_{j}^{k+1} = d_{j}^{-1} x_{j}^{k} \sum_{n=1}^{N} P_{nj} \frac{1}{(P x_{j}^{k})_{n}}
\]
(6.4)
In [26], Parra and Barrett use the original EM formulation of to derive this version of the EMML algorithm applicable to list-mode emission tomographic image reconstruction. The functions \(f_{j}(\cdot)\) in their formulation are pdf. In Huesman et al. [16], the same algorithm is derived starting with a bin-mode formulation and the Poisson model and then rewriting the likelihood in list-mode format.

It follows from (5.4) that the list-mode case limit vector \(x_{\infty}\) has the property \(x_{\infty} = \sum_{i=1}^{I} d_{j} = y_{+} = N\); so, for the solution vector, the expected number of detected events equals the actual number of events on the list.

It is helpful to note that even though the list-mode EM algorithm is derived here as a particular case of an iterative scheme formally identical to a regularization method, that is, the MAP approach of Lange et al., having \(\gamma_{j} = 0\) can mean that an additional regularization term may be needed to achieve a smooth reconstruction. The snowly ML reconstructions one often sees when the EMML algorithm is applied to noisy data can result when the value zero is assigned to many of the voxels in the image, in an apparently random fashion. The images can be improved by regularization methods, such as the MAP method of Lange et al., that prevent zero voxel values; zero values are not prevented in the list-mode EM algorithm, hence noisy reconstructions are still possible.

VII. Bin-Mode Data in List-Mode Form

Suppose now that we have predetermined outcomes, or bins, indexed by \(i = 1, \ldots, I\), and that each datum in the list is simply the index number of the bin associated with that item; that is, \(z_{n} = i_{n} \in \{1, 2, \ldots, I\}\) for each \(n\). Let \(y_{i} > 0\) denote the number of times that the index \(i\) occurs in the list. We assume, as before, that the items in the list were generated independently and that \(N\), the number of items in the list, is Poisson distributed, with mean \(x_{+} d\). We make no further assumptions about the distributions of the counts \(y_{i}\).

In this case, the set \(Z\) of possible events contains the set of bin indexes, but it does not make sense to think of \(Z\) as equal to this set. We must also account for events that are not detected. One way to do this is to let \(Z\) consist of the indexes \(Z = \{1, 2, \ldots, I, I + 1\}\), with the event \(z = I + 1\) corresponding to an undetected event. Then \(\sum_{i=1}^{I} f_{j}(i) = 1\), but \(\sum_{i=1}^{I} f_{j}(i) = s_{j} \leq 1\) for each \(j\). It makes sense now to let \(d_{j} = s_{j}\). Since all we have are bin counts \(y_{i}\), for \(i = 1, 2, \ldots, I\), when we convert the bin-mode data to list-mode, the index \(z = I + 1\) never appears on the list.

The log likelihood function \(LL(x)\) can then be rewritten as
\[
LL(x) = -x_{+} d + N \log(x_{+} d) + \sum_{i=1}^{I} y_{i} \log(P_{i})
\]
(7.1)
where \(P\) is now the matrix with entries \(P_{ij} = f_{j}(i)\) and \((P_{x_{i}}) = \sum_{j=1}^{I} P_{ij} x_{j}\). Then the resulting iterative algorithm is exactly the EMML method. The EMML update equation has been used elsewhere for probabilistic mixtures without the Poisson assumption; see, for example, Lucy [22].

It is also possible to go the other way and to formulate list-mode data in bin-mode form. Typically, this involves loss of information, as distinguishable events are placed in the same bin.

VIII. Block-Iterative Versions of Our Algorithm

In recent years, block-iterative or ordered-subset versions of several algorithms have been used in order to accelerate the production of usable images (see, for example, [8] and the refer-
ences given there, as well as [25]). Such methods are based on
the decomposition of the index set \{n = 1, \ldots, N\} into (not
necessarily disjoint) subsets or blocks, \(B_t, t = 1, \ldots, T\). At
each step of the iterative procedure we select a current block
and we use only those data values \(y_n\) with \(n\) in the current block.
Such methods typically do not converge to a single limit vector,
but instead, have subsequential convergence to a limit cycle of
(generally) \(T\) distinct vectors. Proof of such subsequential
convergence has been given only for the simplest algorithms based
on least squares distance and no proof of subsequential conver-
genence is known for algorithms based on the KL distance.

For \(t = 1, \ldots, T\) and \(j = 1, \ldots, J\), let \(s_{jt} = \sum_{n \in B_t} P_{nj}\)
and let

\[
m_t = \max \{s_{jt} / s_j \mid j = 1, \ldots, J\}.
\]

Our block-iterative version of the algorithm in (5.3) is the
following.

Algorithm 8.1. Let \(x^0\) be an arbitrary vector in \(X\) with
positive entries. For \(k = 0, 1, \ldots\), let \(t(k) = k(\text{mod } T) + 1\). If, for
\(k = 1, 2, \ldots\), we have \(x^k \in X\), let the iterative step be given by

\[
x_j^{k+1} = a_{jk} x_j^k + b_{jk} s_{jt(k)} x_j^k \sum_{n \in B_{t(k)}} \frac{y_n}{(P x^k)_n} + c_j \gamma_j
\]

(8.1)

where

\[
a_{jk} = (s_j - m^{-1}_{t(k)} s_{jt(k)}) / (s_j + \beta_j);
\]

\[
b_j = m^{-1}_{t(k)} s_{jt(k)} / (s_j + \beta_j);
\]

\[
c_j = \beta_j (s_j + \beta_j).
\]

With \(\beta_j = 0\), this algorithm becomes the RBI-EMML method
discussed in [7].

As with other block-iterative methods, the design of the
blocks and the order in which they are used will affect the
degree of acceleration achieved. For a discussion of related
issues, see the recent article by Levkovitz et al. [21].

IX. PROBABILISTIC MIXTURES

We noted with reference to (4.1) that we are using a proba-
bilistic mixture model with unknown mixing coefficients.
To illustrate the notion of a probabilistic mixture problem, we
present the following simple example.

Suppose there are \(110\) possible outcomes from rolling two dice.
The marbles in each bowl come in \(I\) different colors, with \(p_j(i)\) the known proportion of marbles in the
\(i\)th bowl having the \(i\)th color; note that, for each fixed \(i\), we have
\(\sum_{i=1}^I p_j(i) = 1\). A pair of possibly weighted dice is
rolled and one marble is selected from the bowl whose index \(i\)
is the number shown on the dice. The index \(i\) associated with
the color of the selected marble is recorded and then the marble
is returned to the bowl. This experiment is repeated many times.

Consider now the random variable \(Z\), whose value is the
index of the color of the next marble selected. The probability
that \(Z\) will take the value \(i\), call it \(P(i)\), depends both on what the
dice show and what the selected bowl contains. This is a mix-
ture. We have

\[
p(i) = \sum_{j=2}^{12} w_j P_j(i)
\]

(9.1)

where \(w_j\) is the probability that \(j\) will be rolled on the dice.
The mixture problem is to estimate these \(w_j\) on the basis of
observation of the color of the many selected marbles.

Suppose that the experiment is repeated \(N\) times and the list
of (indexes of) observed colors is \(i_1, i_2, \ldots, i_N\). Let \(w\) denote
the column vector with entries \(w_j\). The likelihood function in
this case is then

\[
L(w) = \prod_{n=1}^N p(i_n).
\]

(9.2)

We can rewrite \(L(w)\) as

\[
L(w) = \prod_{i=1}^I y_i \log p(i),
\]

(9.3)

where \(y_i\) denotes the number of times the index \(i\) appears in
the list. Maximizing \(L(w)\) is equivalent to maximizing the log
likelihood, given by

\[
L(w) = \sum_{i=1}^I y_i \log p(i).
\]

(9.4)

The problem then is to maximize

\[
\sum_{i=1}^I y_i \log \left( \sum_{j=2}^{12} w_j P_j(i) \right)
\]

subject to \(\sum_{j=2}^{12} w_j = 1\) and \(w_j \geq 0\) for all \(j\). We can calculate
this maximizer using our iterative algorithm.

We shall consider only those mixture coefficient vectors \(w\) in
the set

\[
W = \left\{ w \mid \sum_{i=1}^I w_j P_j(i) > 0, \quad i = 1, \ldots, I \right\}.
\]

Let \(q_i = y_i / y_+\) for each \(i\).

Conveniently, the solution of the problem, to maximize

\[
\sum_{i=1}^I q_i \log \left( \sum_{j=1}^J w_j P_j(i) \right) - \sum_{j=1}^J w_j P_j(i)
\]

subject only to \(w_j \geq 0, j = 1, \ldots, J\), has the property that
\(\sum_{j=2}^J w_j = 1\), since, for \(j = 1, \ldots, J\), we have \(\sum_{i=1}^I P_j(i) = 1\).
Therefore, we can obtain our desired estimate of \(w\) by per-
forming this maximization, using the algorithm given by (5.3),
with \(i\) in place of \(n, q_i\) in place of \(y_n, \beta_j = 0\) and \(P_j(i) = P_j(i)
\) in place of \(P_{nj}\). The iterative steps become

\[
w_{i+1} = w_i \sum_{i=1}^I \frac{P_j(q_i / P_{w^{k+i}})}{(P_{w^{k+i}})_{i}}.
\]

(9.5)

In the probabilistic mixture problem described by (4.1),
the functions \(f_j(.)\) are assumed known and we need only
estimate the mixing coefficients, \( x_j / x_+ \). More complicated mixtures are often used in image processing and elsewhere. In [28], Samadani employs a mixture model in which \( J = 3 \) and the \( f_j(\cdot) \) have the known form of gamma distributions, but involve unknown parameters, which must also be estimated. His iterative algorithm has two alternating steps, the first of which updates the mixing coefficients using the same method we presented here.

In [10], a somewhat different approach to such complicated mixture problems was presented. Instead of using a small value of \( F \) and a family of functions \( f_f(\cdot) \) involving unknown parameters, one can consider a much larger \( J \) and a family of functions \( f_f(\cdot) \) each associated with a known set of parameter values. Our assumption then is that the mixing coefficients are nonzero for only a small number of \( j \) and the objective is to determine which values of \( j \) these are. High-resolution spectrum analysis methods were modified to provide estimates of the mixing coefficients that are predisposed to single out only a small number of nonzero values.

X. CLOSING REMARKS

The main goal of this paper has been to provide an optimization-theoretic derivation and proof of convergence of the list-mode EM algorithm. We have shown that the list-mode EM algorithm is a special case of a more general iterative method that also includes the EMML method of SPECT and the MAP gamma priors procedure of Lange et al. Our general algorithm is formally an extension of this MAP procedure, in which the parameters are allowed to assume values outside the gamma distribution model. The proof of convergence is also more general, and does not require unique minimizers or any restrictions on the values of \( N \) and \( J \). Since \( \gamma_{j} = 0 \) is permitted, solutions on the boundary are once again possible, unlike in the MAP case.

For the list-mode ET problem, the method presented here is essentially the same as that given by Bouwens et al. [2], [3]. As they point out, we do not need to know the individual functions \( f\_j(\cdot) \) or even their values \( f\_j(z_n) \) at the items in the list; all we need are the values \( f\_j(z_n) \), with \( t_n = \sum_{j=1}^{J} f\_j(z_n) \). The method in [26] does require specification of the densities \( f\_j(\cdot) \), which then requires detailed description of what information is contained in the entries of each list item. In [26], the authors give an analytic expression for \( f\_j(z_n) \) for 2-D PET reconstruction, in which each list item contains two 2-D position coordinates and the time-of-flight. The method here, as modified by Bouwens, et al., requires, for each fixed \( n \), only the specification of a finite vector of probabilities describing the relative likelihoods that an emission at any one of the various pixel locations was responsible for the \( n \)th list item.

One important aspect of the list-mode problem that we have not touched on here is the actual determination of the various probabilities needed. This issue is treated in some detail in [26], as well as in [2] and [3].

APPENDIX

PROOF OF THE THEOREM

We present here a complete, elementary and self-contained proof of convergence of our iterative algorithm, which, as we have seen, contains the EMML method, the gamma prior MAP method and the list-mode EM method as particular cases.

As we saw above, minimizing the function \( G(x) \) given by (4.8) is equivalent to minimizing a related function in which the \( \beta_j \) are replaced by positive quantities. From this rewriting, we conclude, first, that the function \( G \) does attain its minimum value, and second, that the \( x^k \) generated by the algorithm are positive for each \( k \) and \( j \). In the proof that follows, we are assuming only that \( \beta_j > 0 \) for each \( j \).

For any nonnegative vector \( x \) in the set \( X = \{x\vert(Px)_n > 0, n = 1, 2, \ldots, N\} \), let \( r(x) \) and \( q(x) \) denote the \( N \) by \( J \) arrays with entries \( r(x)_nj = x_j P_{nj} y_n / (Px)_n \) and \( q(x)_nj = x_j P_{nj} \), respectively. For any nonnegative vectors \( x \) and \( z \) in \( X \), define

\[
H(x, z) = KL(r(x), q(z)) + \sum_{j=1}^{J} \beta_j KL(\gamma_j, x_j).
\]

For fixed \( x \) in \( X \), denote by \( x^* \) the vector with entries given by

\[
x^*_j = \alpha_j x^*_j - x_j \sum_{n=1}^{N} P_{nj} y_n / (Pz)_n + (1 - \alpha_j) x_j.
\]

From Fact 2 below, we conclude that \( z = x^* \) is the value of \( z \) for which \( H(x, z) \) is minimized. It is easy to see that \( (x^*)^2 = x^{k+1} \) as given by (5.2); it is in this sense that we consider the algorithm to be one of sequential unconstrained minimization [15]. Note also that \( KL(r(x), q(x)) = KL(y, Px) \) and \( H(x, x) \) is \( G(x) \).

The following useful identities are established by tedious, but elementary, calculations, which we omit.

**Fact 1:**

\[
KL(r(x), q(z)) = KL(r(z), q(z)) + KL(r(x), r(z)).
\]

**Fact 2:**

\[
H(x, z) = H(x, x^*) + \sum_{j=1}^{J} (s_j + \beta_j) KL(x^*_j, z_j).
\]

From (5.4), we obtain

\[
\sum_{j=1}^{J} (s_j + \beta_j) x^{k+1}_j = \sum_{n=1}^{N} y_n + \sum_{j=1}^{J} \beta_j \gamma_j
\]

from which it follows that the sequence \( \{x^{k}\} \) is bounded. Let \( x^* \) denote an arbitrary cluster point of the sequence \( \{x^{k}\} \).

Using the previous two facts, we obtain

**Fact 3:**

\[
G(x^{k}) = G(x^{k+1}) + KL(r(x^{k+1}), r(x^{k+1}))) + \sum_{j=1}^{J} (s_j + \beta_j) KL(x^{k+1}_j, x^{k}_j).
\]

From Fact 3, it follows that the sequence \( \{G(x^{k})\} \) is decreasing and the sequence \( \sum (s_j + \beta_j) KL(x^{k+1}_j, x^{k}_j) \) converges to zero; consequently \( G(x^*) = +\infty \), so \( x^* \) is in \( X \), and \( x^* \) is a fixed point of the iteration, that is, \( x^* = x^* \).

Now let \( \hat{x} \) be a nonnegative minimizer of the function \( G(x) \); since \( G(\hat{x}) < +\infty \), we know \( \hat{x} \) is in \( X \). It follows from Fact
2 that  is also a fixed point of the iteration. The proof of convergence is based on using the facts above to write the quantity $H(\hat{x}, x^k)$ in two equivalent ways. Specifically, we have

$$H(\hat{x}, x^k) = G(\hat{x}) + \sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^k)$$  \hspace{1cm} (10.3)$$

and

$$H(\hat{x}, x^k) = F(x^k) + KL(\hat{x}, x^k)$$  \hspace{1cm} (10.4)$$

Since $G(x^k) \geq G(\hat{x})$ it follows that

$$\sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^k) \geq KL(\hat{x}, x^k)$$  \hspace{1cm} (10.5)$$

The key step in the proof is to obtain the following inequality:

$$KL(\hat{x}, x^k) \geq \sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^{k+1})$$  \hspace{1cm} (10.6)$$

Combining inequalities (10.5) with (10.6) will produce what we have called elsewhere the double inequality on which the proof rests. Inequality (10.6) appears in various guises in other papers; in Csiszár and Tusnády [12] it is their four-point property, while in Iusem [17] it is a consequence of the joint convexity of the KL distance. Indeed, it has been shown by Eggert and LaRiccia that joint convexity of more general Bregman distances implies the four-point property.

The next two lemmas concerning the KL distance are helpful at this stage.

**Lemma 10.1:** For any nonnegative vectors $x$ and $z$, with $x_+ = \sum_{j=1}^n x_j$ and $z_+ = \sum_{j=1}^n z_j > 0$, we have

$$KL(x, z) = KL(x_+, z_+) + KL(x, (x_+/z_+)z).$$

**Proof:** The proof is an easy calculation.

**Lemma 10.2:** For any positive scalars $a$, $b$, and $c$, with $a \neq b$, we have

$$KL(a, b) > KL(a + c, b + c).$$

**Proof:** With $a$ and $b$ fixed, the function $g(c) = KL(a + c, b + c)$ has derivative $g'(c) = 1 - t + \log t < 0$, for $t = (a + c)/(b + c)$, so $g(c)$ is strictly decreasing for $c \geq 0$.

Applying the first lemma to the sum over the index $n$, we obtain

$$KL(\hat{x}, x^k) \geq \sum_{j=1}^J KL(\hat{x}_j, x_j^k).$$  \hspace{1cm} (10.7)$$

This can be rewritten as

$$KL(\hat{x}, x^k) \geq \sum_{j=1}^J KL((s_j + \beta_j)^2 \hat{x}_j - \beta_j \gamma_j, (s_j + \beta_j)^2 x_j^{k+1} - \beta_j \gamma_j).$$  \hspace{1cm} (10.8)$$

Applying the second lemma and using the fact that $b_j \gamma_j \geq 0$, we get the inequality (10.6).

Combining inequalities (10.5) and (10.6), we conclude that the sequence

$$\sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^*)$$

is decreasing and that $KL(\hat{x}_j, x_j^*) < +\infty$ for each $j$. Therefore, $\hat{x}_j > 0$ implies $x_j^* > 0$. Also

$$\sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^*) = KL(\hat{x}, x^*) < +\infty.$$ 

We then have

$$H(\hat{x}, x^*) = G(\hat{x}) + \sum_{j=1}^J (s_j + \beta_j)KL(\hat{x}_j, x_j^*)$$  \hspace{1cm} (10.9)$$

from (10.3) and

$$H(\hat{x}, x^*) = G(x^*) + KL(\hat{x}, x^*)$$  \hspace{1cm} (10.10)$$

from (10.4), from which we conclude that $G(\hat{x}) = G(x^*)$. Therefore, $x^*$ is a nonnegative minimizer of $G(x)$ and we can replace the generic $x$ with $x^*$ in the results just obtained. In particular, we now know that the sequence

$$\sum_{j=1}^J (s_j + \beta_j)KL(x_j^*, x_j^*)$$

is decreasing; but since a subsequence converges to zero, the entire sequence converges to zero, which tells us that $x^k \to x^*$. This concludes the proof of the theorem.

**ACKNOWLEDGMENT**

The author wishes to thank L. Bouwens, Ghent University, Belgium, who supplied references [2] and [3], and S. Glick and M. King of the University of Massachusetts Medical School for several helpful discussions. The author also thanks an anonymous referee for the idea of rewriting the function $G(x)$ in terms of positive parameters.

The contents of this paper are solely the responsibility of the author and do not necessarily represent the official views of NIH.

**REFERENCES**


