Globally Convergent Algorithms for Maximum a Posteriori Transmission Tomography

Kenneth Lange and Jeffrey A. Fessler
The University of Michigan

Email: klange@ucla.edu fessler@umich.edu

IEEE T-IP 4(10):1430-8, Oct. 1995

ABSTRACT

This paper reviews and compares three maximum likelihood algorithms for transmission tomography. One of these algorithms is the EM algorithm, one is based on a convexity argument devised by De Pierro in the context of emission tomography, and one is an ad hoc gradient algorithm. The algorithms enjoy desirable local and global convergence properties and combine gracefully with Bayesian smoothing priors. Preliminary numerical testing of the algorithms on simulated data suggest that the convex algorithm and the ad hoc gradient algorithm are computationally superior to the EM algorithm. This superiority stems from the larger number of exponentiations required by the EM algorithm. The convex and gradient algorithms are well adapted to parallel computing.

Key words: maximum likelihood, smoothing prior, EM algorithm, convergence

I. INTRODUCTION

THE value of the EM algorithm in emission tomography is now well established [17, 22, 24]. Not as widely appreciated is the potential of the EM algorithm in transmission tomography [17]. This paper reviews the EM algorithm for transmission tomography and compares it to two algorithms recently introduced by Lange et al. [16] and Lange [12].

The traditional method of image reconstruction in transmission tomography relies on Fourier analysis and the Radon transform [10]. An alternative to this deterministic reconstruction method is to pose an explicitly stochastic model that permits parameter estimation by maximum likelihood [17]. In this context the EM algorithm provides an easily implemented method for searching the likelihood surface. This does not mean that the EM or competing stochastic algorithms can match Fourier methods in computational speed. But the increased realism possible with a stochastic model does promise better image reconstruction with lower patient radiation dose.

The object of transmission tomography is to reconstruct the local attenuation properties of the object being imaged. Attenuation is roughly to be equated with density. In an imaging experiment, X-rays or γ-rays are beamed from an external source through the imaged object. These high energy photons can be stopped or deflected by the object, or they can be detected by a device on the opposite side of the object. Only a fraction of the photons successfully travel from source to detector along a given flight path (projection). The probability of a photon escaping attenuation along a projection is given by exponentiating the negative of the line integral of the attenuation density along the projection. In deterministic reconstruction, these line integrals are mathematically massaged to given the final image. No account is taken of the fact that the observed data actually consist of photon counts.

The stochastic model depends on dividing the object of interest into small non-overlapping regions of constant attenuation called pixels. Typically the pixels are squares. To each pixel is assigned an attenuation parameter. In the absence of the intervening object, the number of photons generated and ultimately detected along a projection follows a Poisson distribution. Attenuation randomly thins these photons. Since thinning a Poisson process yields a Poisson process, the number of photons detected also follows a Poisson distribution. The detected photon counts constitute the observed data for stochastic reconstruction.

The remainder of this paper builds on the above verbal model of transmission tomography. Section 2 motivates three competing algorithms for maximum likelihood estimation of the attenuation parameters. Local convergence of the algorithms is examined under the simplifying assumption that the maximum point is interior to the feasible region. Section 3 outlines how the algorithms can be amended to incorporate Bayesian smoothing parameters. Section 4 proves that two of the algorithms are globally convergent. Section 5 compares the numerical performance of the algorithms on simulated data. The concluding discussion in Section 6 draws some preliminary conclusions about the numerical efficiency of the algorithms and suggests topics for further research.

II. ALGORITHMS FOR TRANSMISSION TOMOGRAPHY

The parameters of interest in transmission tomography are the linear attenuation coefficients μ_j defined for each pixel j. Since μ_j is the probability of photon capture per unit length of

Research supported in part by USPHS Grants CA-16042[1] and CA-60711[2] and DOE Grant DE-FG02-87ER60561[2].

pixel j, we have the obvious physical constraint $\mu_j \geq 0$. The Poisson nature of X-ray generation implies that the various projections are independent and that the loglikelihood of the observed photon counts Y_i can be written as

$$L(\mu) = \sum_i \{-d_i e^{-\langle l_i, \mu \rangle} - Y_i \langle l_i, \mu \rangle\} + c. \tag{1}$$

In equation (1), d_i is the expected number of photon counts leaving the source along the ith projection; c is an irrelevant constant; μ is the vector of attenuation parameters μ_j; l_i is the vector of intersection lengths l_{ij} for the ith projection; and $\langle l_i, \mu \rangle$ denotes the inner product $\sum_j l_{ij}\mu_j$. This inner product can be interpreted as the line integral of the discretized attenuation from source to detector along projection i.

A. EM Algorithm

One can deduce an EM algorithm for this model by defining the complete data as the number of photons entering and leaving each pixel along each projection. Let U_{ij} and V_{ij} be the numbers of photons entering and leaving, respectively, pixel j along projection i. The E step of the EM algorithm requires the conditional expectations $M_{ij} = E(U_{ij} \mid Y_i, \mu^n)$ and $N_{ij} = E(V_{ij} \mid Y_i, \mu^n)$; Lange and Carson [17] prove that

$$\begin{aligned} M_{ij} &= Y_i + d_i e^{-\sum_{k \in S_{ij}} l_{ik}\mu_k^n} - d_i e^{-\langle l_i, \mu^n \rangle} \\ N_{ij} &= Y_i + d_i e^{-\sum_{k \in S_{ij} \cup \{j\}} l_{ik}\mu_k^n} - d_i e^{-\langle l_i, \mu^n \rangle}, \end{aligned}$$

where S_{ij} is the set of pixels between the source and pixel j along projection i. The $Q(\mu \mid \mu^n)$ function of the EM algorithm [1] then turns out to be

$$\begin{aligned} & Q(\mu \mid \mu^n) \\ = & \sum_i \sum_j \left[-N_{ij} l_{ij} \mu_j + (M_{ij} - N_{ij}) \ln(1 - e^{-l_{ij}\mu_j}) \right]. \end{aligned}$$

The M step of the EM algorithm consists in maximizing $Q(\mu \mid \mu^n)$ with respect to μ. Setting the partial derivative of $Q(\mu \mid \mu^n)$ with respect to μ_j equal to 0 yields the transcendental equation

$$0 = \sum_i -N_{ij} l_{ij} + \sum_i \frac{(M_{ij} - N_{ij}) l_{ij}}{e^{l_{ij}\mu_j} - 1}. \tag{2}$$

Lange and Carson [17] are quick to point out that the solution of this transcendental equation can be approximated by

$$\mu_j^{n+1} = \frac{\sum_i (M_{ij} - N_{ij})}{\frac{1}{2}\sum_i (M_{ij} + N_{ij}) l_{ij}}, \tag{3}$$

assuming the product $l_{ij}\mu_j^{n+1}$ is small. Ollinger [19] argues that it is safer to solve (2) iteratively by Newton's method or like algorithms.

B. Gradient Algorithm

The EM algorithm is cumbersome because of the large number of exponentiations it entails. An alternative algorithm suggested in Lange et al. [16] updates the attenuation parameter vector μ by

$$\begin{aligned} \mu_j^{n+1} &= \mu_j^n \frac{\sum_i d_i e^{-\langle l_i, \mu^n \rangle} l_{ij}}{\sum_i Y_i l_{ij}} \\ &= \mu_j^n + \frac{\mu_j^n}{\sum_i Y_i l_{ij}} \frac{\partial}{\partial \mu_j} L(\mu^n). \end{aligned} \tag{4}$$

This is a scaled gradient algorithm with a nonconstant diagonal scaling matrix. For brevity we will refer to (4) simply as a gradient algorithm. It can be heuristically motivated by noting that $d_i e^{-\langle l_i, \mu^n \rangle}$ is the expected number of photons detected along projection i. Y_i is the observed number of photons detected. Each of these is weighted by the intersection length l_{ij} for pixel j, and the results are summed over all projections i intersecting pixel j. If μ_j^n is too large, the numerator tends to be smaller than the denominator in (4) and $\mu_j^{n+1} < \mu_j^n$. If μ_j^n is too small, the reverse $\mu_j^{n+1} > \mu_j^n$ tends to occur. Unfortunately, there are no obvious guarantees that the algorithm will either increase the loglikelihood $L(\mu)$ or preserve parameter nonnegativity constraints. These defects can be remedied by taking only a fractional step in the direction implied by the increment $\mu^{n+1} - \mu^n$ defined in (4).

C. Convex Algorithm

Lange [12] discusses yet a third algorithm for transmission tomography. This algorithm bears a striking resemblance to the EM algorithm although it does not invoke any notions of missing data. To motivate the algorithm, rewrite the loglikelihood as

$$L(\mu) = -\sum_i f_i(\langle l_i, \mu \rangle)$$

using the strictly convex functions $f_i(t) = d_i e^{-t} + Y_i t$. We can construct the algorithm by imitating certain arguments of De Pierro for emission tomography [2, 3]. The crux of the matter is that at iteration n

$$\begin{aligned} L(\mu) &= -\sum_i f_i\Big(\sum_j \frac{l_{ij}\mu_j^n}{\langle l_i, \mu^n \rangle} \frac{\mu_j}{\mu_j^n} \langle l_i, \mu^n \rangle\Big) \\ &\geq -\sum_i \sum_j \frac{l_{ij}\mu_j^n}{\langle l_i, \mu^n \rangle} f_i\Big(\frac{\mu_j}{\mu_j^n} \langle l_i, \mu^n \rangle\Big) \\ &= Q(\mu \mid \mu^n), \end{aligned} \tag{5}$$

with strict inequality unless $\frac{\mu_j}{\mu_j^n}\langle l_i, \mu^n \rangle = \frac{\mu_k}{\mu_k^n}\langle l_i, \mu^n \rangle$ for all i and all $j \neq k$. If $\mu_j = \mu_j^n$ for all j, then the inequality (5) is an equality. The function $Q(\mu \mid \mu^n)$ defined on the right of (5) is the analog of the function by the same name in classical EM theory [1]. It is specifically designed so that the difference $L(\mu) - Q(\mu \mid \mu^n)$ attains its minimum of 0 at $\mu = \mu^n$.

Just as in the usual EM theory, we choose μ^{n+1} to maximize $Q(\mu \mid \mu^n)$. If μ^{n+1} is so selected, then

$$\begin{aligned} L(\mu^{n+1}) &= L(\mu^{n+1}) - Q(\mu^{n+1} \mid \mu^n) + Q(\mu^{n+1} \mid \mu^n) \\ &\geq L(\mu^n) - Q(\mu^n \mid \mu^n) + Q(\mu^n \mid \mu^n) \\ &= L(\mu^n), \end{aligned}$$

with strict inequality when $\mu^{n+1} \neq \mu^n$. We will refer to this method of selecting μ^{n+1} as the convex algorithm.

To maximize $Q(\mu \mid \mu^n)$ set

$$\begin{aligned} 0 &= \frac{\partial}{\partial \mu_j} Q(\mu \mid \mu^n) \\ &= -\sum_i l_{ij} f_i'(\frac{\mu_j}{\mu_j^n}\langle l_i, \mu^n\rangle) \\ &= -\sum_i l_{ij}[-d_i e^{-\frac{\mu_j}{\mu_j^n}\langle l_i,\mu^n\rangle} + Y_i]. \end{aligned} \tag{6}$$

The transcendental equation (6) can not be solved exactly. It does have a unique solution, however. Ordinarily, this solution is positive. Indeed, the right hand side of (6) is strictly decreasing in μ_j. For $\mu_j = 0$, its value is $-\sum_i l_{ij}[-d_i + Y_i]$, which is usually positive because $Y_i \approx d_i$ for a projection that does not sample the object and $Y_i \ll d_i$ for a projection that does substantially sample the object. For $\mu_j = \infty$, the right hand side of (6) is $-\sum_i l_{ij} Y_i$, which is negative. Thus typically the solution falls somewhere on the open interval $(0, \infty)$.

We can solve equation (6) by Newton's method. Since

$$\begin{aligned} &\frac{\partial^2}{\partial \mu_j^2} Q(\mu \mid \mu^n) \mid_{\mu=\mu^n} \\ &= -\sum_i \frac{l_{ij}}{\mu_j^n}\langle l_i, \mu^n\rangle f_i''(\frac{\mu_j}{\mu_j^n}\langle l_i, \mu^n\rangle) \\ &= -\sum_i \frac{l_{ij}}{\mu_j^n}\langle l_i, \mu^n\rangle d_i e^{-\frac{\mu_j}{\mu_j^n}\langle l_i,\mu^n\rangle}, \end{aligned}$$

and

$$\frac{\partial}{\partial \mu_j} Q(\mu \mid \mu^n) \mid_{\mu=\mu^n} = \frac{\partial}{\partial \mu_j} L(\mu^n)$$

for $\mu_j^n > 0$, one step of Newton's method gives the approximate solution

$$\begin{aligned} \mu_j^{n+1} &= \mu_j^n + \frac{\mu_j^n}{\sum_i l_{ij}\langle l_i, \mu^n\rangle d_i e^{-\langle l_i,\mu^n\rangle}} \frac{\partial}{\partial \mu_j} L(\mu^n) \\ &= \mu_j^n + \frac{\mu_j^n \sum_i l_{ij}[d_i e^{-\langle l_i,\mu^n\rangle} - Y_i]}{\sum_i l_{ij}\langle l_i, \mu^n\rangle d_i e^{-\langle l_i,\mu^n\rangle}} \\ &= \mu_j^n \frac{\sum_i l_{ij}[d_i e^{-\langle l_i,\mu^n\rangle}(1 + \langle l_i, \mu^n\rangle) - Y_i]}{\sum_i l_{ij}\langle l_i, \mu^n\rangle d_i e^{-\langle l_i,\mu^n\rangle}}. \end{aligned} \tag{7}$$

This approximate solution of the M step coincides with the algorithm proposed in equation (9) of [12]. The idea of solving the M step approximately by one step of Newton's method is motivated in [14]. One of the results in [14] says that even this approximate solution of the M step leads to an increase in $L(\mu)$ in a neighborhood of the optimal point. (This theory does not quite fit the current problem because of the presence of the boundaries $\mu_j \geq 0$.)

The algorithm (7) also has the potential disadvantage of giving $\mu_j^{n+1} < 0$ when $\mu_j^n > 0$. This drawback is apt to be more theoretical than practical, however. As argued above, the exact solution of (6) is usually positive. If the Newton iterate (7) approximates this solution well, then the Newton iterate will usually be positive as well.

D. Local Convergence

To analyze the behavior of the algorithm (7) in a neighborhood of the maximum point $\hat{\mu}$, we make the simplifying assumptions that $\hat{\mu}$ exists, is unique, and occurs in the interior of the feasible region. We can then view the iterates given by (7) as moving toward a fixed point of the map

$$G(\mu) = \mu + D(\mu) dL(\mu),$$

where $D(\mu)$ is the diagonal matrix with jth diagonal entry

$$D_{jj}(\mu) = \frac{\mu_j}{\sum_i l_{ij}\langle l_i, \mu\rangle d_i e^{-\langle l_i,\mu\rangle}},$$

and $dL(\mu)$ is the score vector with jth entry

$$\frac{\partial}{\partial \mu_j} L(\mu) = \sum_i l_{ij}[d_i e^{-\langle l_i,\mu\rangle} - Y_i].$$

According to a theorem of Ostrowski [20], the fixed point $\hat{\mu}$ is locally attractive provided the spectral radius of the differential $dG(\hat{\mu})$ is strictly less than 1. This spectral radius determines the linear convergence rate of the algorithm. Since $dL(\hat{\mu}) = 0$, it follows that

$$\begin{aligned} dG(\hat{\mu}) &= I + D(\hat{\mu}) d^2 L(\hat{\mu}) \\ &= D(\hat{\mu})[D(\hat{\mu})^{-1} + d^2 L(\hat{\mu})], \end{aligned}$$

where $d^2 L(\hat{\mu})$ is the second differential or Hessian matrix of $L(\mu)$. To estimate the spectral radius of $dG(\hat{\mu})$ requires a lemma.

Lemma 1 *Suppose A and B are symmetric matrices with A and B positive definite and $A - B$ positive semidefinite. Then the eigenvalues of $A^{-1}(A - B)$ lie on $[0, 1)$.*

PROOF: This well-known result is proved with minor notational differences by Green [9]. ∎

In the usual EM theory [1], the matrix difference $A - B$ is identified with the expected information of the complete data given the observed data. In the current algorithm, we identify A with $D(\hat{\mu})^{-1}$ and B with $-d^2 L(\hat{\mu})$. Assuming that all $\hat{\mu}_j > 0$, the matrix A is positive definite. Positive definiteness of B is a consequence of strict concavity of $L(\mu)$. Strict concavity is hard to verify in practice; a necessary condition is that the number of projections exceeds the number of pixels.

In any case to apply the lemma, we need to verify that $A - B$ is positive semidefinite. Direct computation with an arbitrary vector v gives

$$\begin{aligned} v^t(A - B)v &= \sum_i d_i e^{-\langle l_i,\hat{\mu}\rangle}\langle l_i, \hat{\mu}\rangle \sum_j l_{ij}\frac{v_j^2}{\hat{\mu}_j} \\ &\quad - \sum_i d_i e^{-\langle l_i,\hat{\mu}\rangle}\langle l_i, v\rangle^2. \end{aligned} \tag{8}$$

Now Cauchy's inequality implies

$$\langle l_i, \hat{\mu}\rangle[\sum_j l_{ij}\frac{v_j^2}{\hat{\mu}_j}] \geq \langle l_i, v\rangle^2. \tag{9}$$

When (9) is multiplied by $d_i e^{-\langle l_i, \hat{\mu} \rangle}$, and the result summed on i, the required inequality

$$v^t(A - B)v \geq 0$$

follows. Local attractiveness is now established by appealing to Ostrowski's theorem and the lemma.

In the event that $\langle l_i, \hat{\mu} \rangle \leq 1$ for all i, the algorithm (4) is also locally attractive. Indeed, the nonnegativity of (9) then yields

$$\begin{aligned} & \sum_i d_i e^{-\langle l_i, \hat{\mu} \rangle} \sum_j l_{ij} \frac{v_j^2}{\hat{\mu}_j} \\ & - \sum_i d_i e^{-\langle l_i, \hat{\mu} \rangle} \langle l_i, v \rangle^2 \quad \geq \quad 0. \end{aligned} \tag{10}$$

Now substitute the identity $\sum_i d_i e^{-\langle l_i, \hat{\mu} \rangle} l_{ij} = \sum_i Y_i l_{ij}$, which follows from $\frac{\partial L}{\partial \mu_j}(\hat{\mu}) = 0$, in (10). This substitution proves the positive semidefiniteness of $A - B$, where $B = -d^2 L(\hat{\mu})$, $A = F(\hat{\mu})^{-1}$, and $F(\hat{\mu})$ is diagonal with jth diagonal entry

$$F(\hat{\mu})_{jj} = \frac{\hat{\mu}_j}{\sum_i Y_i l_{ij}}.$$

Thus, if attenuation is sufficiently weak for $\langle l_i, \hat{\mu} \rangle \leq 1$ to hold uniformly in i, then the gradient algorithm (4) is locally attracted to $\hat{\mu}$.

In practice, the assumption that $\langle l_i, \hat{\mu} \rangle \leq 1$ holds uniformly in i is suspect. If we replace this condition by $\langle l_i, \hat{\mu} \rangle \leq c$ uniformly in i for $c \geq 1$, then the above argument can be amended to show that the gradient algorithm $\mu_j^{n+1} = \mu_j^n + \frac{\mu_j^n}{c \sum_i Y_i l_{ij}} \frac{\partial}{\partial \mu_j} L(\mu^n)$ converges locally.

III. Incorporation of Smoothing Priors

How can the above algorithms be modified to take into account a smoothing prior [6, 7]? The loglikelihood is changed to the log posterior $\Delta(\mu) = L(\mu) - U(\mu)$, where $U(\mu)$ is some energy function penalizing large deviations between neighboring pixels. For the EM algorithm and the convex algorithm, the $Q(\mu \mid \mu^n)$ function is then changed to $Q(\mu \mid \mu^n) - U(\mu)$. The maximum μ^{n+1} of this amended function satisfies

$$0 \quad = \quad \frac{\partial}{\partial \mu_j} Q(\mu \mid \mu^n) - \frac{\partial}{\partial \mu_j} U(\mu), \tag{11}$$

Green [8,9] decouples and approximately solves the set of equations (11) by pretending that the argument of $\frac{\partial}{\partial \mu_j} U(\mu)$ is the constant μ^n instead of the unknown μ.

The Gibbs priors introduced by Geman and McClure [6, 7] take the form

$$U(\mu) \quad = \quad \gamma \sum_{\{j,k\} \in N} w_{jk} \psi(\mu_j - \mu_k),$$

where γ and the weights w_{jk} are positive constants, N is a set of unordered pairs $\{j, k\}$ defining a neighborhood system, and $\psi(r)$, r real, is a potential function. For instance, if the pixels are squares, we might define the weights by $w_{jk} = 1$ for orthogonal nearest neighbors and $w_{jk} = \frac{1}{\sqrt{2}}$ for diagonal nearest neighbors. Defining the pixels as regular hexagons eliminates diagonal nearest neighbors and permits all weights to be equal. The constant γ scales the overall strength assigned to the prior.

Choice of the potential function $\psi(r)$ is the most crucial feature of the Gibbs prior. It is convenient to assume that $\psi(r)$ is even, twice continuously differentiable, and strictly convex with $\psi''(r) > 0$ for all r. Strict convexity leads to strict concavity of the log posterior $\Delta(\mu) = L(\mu) - U(\mu)$ and permits simple modification of the EM algorithm and the convex algorithm. There are many potential functions satisfying these conditions. One obvious example is $\psi(r) = r^2$. This choice tends to deter the formation of boundaries, and Green [8, 9] has suggested the gentler alternative $\psi(r) = \ln[\cosh(r)]$, which grows for large $|r|$ linearly rather than quadratically. Lange [13] lists a number of other potential functions exhibiting linear growth at $|r| = \infty$.

De Pierro [2,3] has proposed an elegant alternative to Green's method of handling the energy function $U(\mu)$ when maximizing $Q(\mu \mid \mu^n) - U(\mu)$. Paralleling his treatment of the loglikelihood, De Pierro exploits convexity so as to reduce maximization of $Q(\mu \mid \mu^n) - U(\mu)$ to a sequence of one-dimensional maximization problems. Now convexity and evenness of the potential function $\psi(r)$ together imply

$$\begin{aligned} & \psi(\mu_j - \mu_k) \\ = \quad & \psi(\frac{1}{2}[2\mu_j - \mu_j^n - \mu_k^n] + \frac{1}{2}[-2\mu_k + \mu_j^n + \mu_k^n]) \\ \leq \quad & \frac{1}{2}\psi(2\mu_j - \mu_j^n - \mu_k^n) + \frac{1}{2}\psi(2\mu_k - \mu_j^n - \mu_k^n), \end{aligned} \tag{12}$$

with strict inequality unless $\mu_j + \mu_k = \mu_j^n + \mu_k^n$. Inequality (12) in turn yields

$$\begin{aligned} & -U(\mu) \\ = \quad & -\gamma \sum_{\{j,k\} \in N} w_{jk} \psi(\mu_j - \mu_k) \\ \geq \quad & -\frac{\gamma}{2} \sum_{\{j,k\} \in N} w_{jk} \psi(2\mu_j - \mu_j^n - \mu_k^n) \\ & -\frac{\gamma}{2} \sum_{\{j,k\} \in N} w_{jk} \psi(2\mu_k - \mu_j^n - \mu_k^n) \\ = \quad & -V(\mu \mid \mu^n). \end{aligned}$$

In both the EM and the convex algorithms, we now substitute the comparison function

$$\Upsilon(\mu \mid \mu^n) \quad = \quad Q(\mu \mid \mu^n) - V(\mu \mid \mu^n)$$

for the comparison function $Q(\mu \mid \mu^n) - U(\mu)$. By construction this amended strictly concave comparison function provides the bound

$$\Delta(\mu) - \Upsilon(\mu \mid \mu^n) \geq \Delta(\mu^n) - \Upsilon(\mu^n \mid \mu^n), \tag{13}$$

on the log posterior $\Delta(\mu)$. If the maximum of $\Upsilon(\mu \mid \mu^n)$ occurs at $\hat{\mu}^n$, then some components $\hat{\mu}_j^n$ may satisfy $\hat{\mu}_j^n = 0$. We can avoid these boundary problems by defining the next iterate μ^{n+1} to have components $\mu_j^{n+1} = \max(\hat{\mu}_j^n, \epsilon \mu_j^n)$ for some

constant ϵ in the open interval $(0,1)$. To prove the crucial inequality $\Delta(\mu^{n+1}) \geq \Delta(\mu^n)$, we now argue as follows. The choice of $\hat{\mu}_j^n$ entails

$$\frac{\partial}{\partial \mu_j}\Upsilon(\mu \mid \mu^n) \mid_{\mu_j=\mu_j^n} (\hat{\mu}_j^n - \mu_j^n) \quad \geq \quad 0 \tag{14}$$

because $\Upsilon(\mu \mid \mu^n)$ separates the parameters μ_j and $\frac{\partial}{\partial \mu_j}\Upsilon(\mu \mid \mu^n)$ has the same sign as the difference $\hat{\mu}_j^n - \mu_j^n$. The inequality (14) remains valid when μ_j^{n+1} is substituted for $\hat{\mu}_j^n$ and the partial derivative is evaluated at any point $\bar{\mu}_j^n$ between μ_j^n and μ_j^{n+1}. Inequality (13) and the mean value theorem then imply

$$\begin{aligned} & \Delta(\mu^{n+1}) \\ \geq \quad & \Delta(\mu^n) + \Upsilon(\mu^{n+1} \mid \mu^n) - \Upsilon(\mu^n \mid \mu^n) \\ = \quad & \Delta(\mu^n) + \sum_j \frac{\partial}{\partial \mu_j}\Upsilon(\mu \mid \mu^n) \mid_{\mu_j=\bar{\mu}_j^n} (\mu_j^{n+1} - \mu_j^n) \\ \geq \quad & \Delta(\mu^n), \end{aligned} \tag{15}$$

with strict inequality when $\mu^{n+1} \neq \mu^n$.

In practice, instead of maximizing $\Upsilon(\mu \mid \mu^n)$, one could settle for one step of Newton's method and use the algorithm

$$\begin{aligned} \mu_j^{n+1} \quad &= \quad \mu_j^n - \frac{\frac{\partial}{\partial \mu_j}\Upsilon(\mu \mid \mu^n) \mid_{\mu=\mu^n}}{\frac{\partial^2}{\partial \mu_j^2}\Upsilon(\mu \mid \mu^n) \mid_{\mu=\mu^n}} \\ &= \quad \mu_j^n - \frac{\frac{\partial}{\partial \mu_j}\Delta(\mu^n)}{\frac{\partial^2}{\partial \mu_j^2}\Upsilon(\mu \mid \mu^n) \mid_{\mu=\mu^n}}. \end{aligned} \tag{16}$$

How to accommodate a smoothing prior in algorithm (4) is not altogether obvious. The problem is that algorithm (4) is not motivated by optimization of a simple function $Q(\mu \mid \mu^n)$ designed to force an increase in $L(\mu)$. It is interesting that the quadratic function

$$\begin{aligned} & Q(\mu \mid \mu^n) \\ = \quad & \sum_j \{\mu_j \sum_i d_i l_{ij} e^{-\langle l_i, \mu^n \rangle} - \frac{(\mu_j)^2}{2} \frac{\sum_i Y_i l_{ij}}{\mu_j^n}\} \end{aligned} \tag{17}$$

is maximized by (4), but this choice of $Q(\mu \mid \mu^n)$ may not guarantee the increase $L(\mu^{n+1}) > L(\mu^n)$. For this reason there is little point in applying De Pierro's transformation of $U(\mu)$ to $V(\mu \mid \mu^n)$. Our limited numerical experience suggests that the alternative algorithm

$$\begin{aligned} \mu_j^{n+1} \quad &= \quad \mu_j^n - \frac{\frac{\partial}{\partial \mu_j}\Delta(\mu^n)}{\frac{\partial^2}{\partial \mu_j^2}Q(\mu \mid \mu^n) \mid_{\mu=\mu^n} - \frac{\partial^2}{\partial \mu_j^2}U(\mu^n)} \\ &= \quad \mu_j^n + \frac{\mu_j^n}{\sum_i Y_i l_{ij} + \mu_j^n \frac{\partial^2}{\partial \mu_j^2}U(\mu^n)} \frac{\partial}{\partial \mu_j}\Delta(\mu^n) \end{aligned} \tag{18}$$

performs better, where $Q(\mu \mid \mu^n)$ is defined by formula (17). This is just one step of Newton's method applied to the function $Q(\mu \mid \mu^n) - U(\mu)$, but omitting the off-diagonal entries of the Hessian $d^2U(\mu^n)$.

IV. Global Convergence of the Algorithms

Both the EM algorithm and the convex algorithm converge to the global maximum of the log posterior. Our proof of this fact incorporates features from previous proofs of Lange and Carson [17] and De Pierro [2]. As noted above, the next iterate μ^{n+1} of either of these algorithms is defined componentwise by $\mu_j^{n+1} = \max(\hat{\mu}_j^n, \epsilon\mu_j^n)$, where $\hat{\mu}_j^n$ either equals 0 or provides the unique root of $\frac{\partial}{\partial \mu_j}\Upsilon(\mu \mid \mu^n) = 0$, and where ϵ is some constant in the interval $(0,1)$. Observe that our definition of μ^{n+1} differs slightly from De Pierro's [2], who takes $\mu_j^{n+1} = \hat{\mu}_j^n$ whenever $\hat{\mu}_j^n > 0$. It is convenient to assume that $\mu_j^0 > 0$ for all j since then $\mu_j^{n+1} > 0$ for all n and j. It is also natural to assume that for each pixel j there is some projection i with $Y_i l_{ij} > 0$.

Convergence of the iterates μ^n hinges on strict concavity of the log posterior. To establish this fact, we assume that the neighborhood system N of the Gibbs prior $U(\mu)$ is connected. If the pixels are considered as nodes of a graph, with two neighboring pixels connected by an edge, then this assumption means that it is possible to find some sequence of edges leading from any pixel to any other pixel. Strict concavity and related properties of the log posterior are summarized in the next lemma. Recall that $\psi(r)$ is even, twice continuously differentiable, and satisfies $\psi''(r) > 0$ for all r.

Lemma 2 *Let $\Delta(\mu) = L(\mu) - U(\mu)$ be the log posterior. Then*

a) $\Delta(\mu)$ is strictly concave.

b) $\lim_{\|\mu\| \to \infty} \Delta(\mu) = -\infty$. Consequently, $\Delta(\mu)$ has a unique maximum.

c) The set $\{\mu : \mu_j \frac{\partial}{\partial \mu_j}\Delta(\mu) = 0 \text{ for all } j\}$ of stationary points of $\Delta(\mu)$ is finite.

PROOF: Strict concavity is verified by examining the quadratic form

$$\sum_j \sum_k \xi_j \frac{\partial^2}{\partial \mu_j \partial \mu_k}\Delta(\mu)\xi_k,$$

which reduces to

$$\begin{aligned} & -\sum_i d_i e^{-\langle l_i, \mu \rangle} \langle l_i, \xi \rangle^2 \\ & -\gamma \sum_{\{j,k\} \in N} w_{jk}\psi''(\mu_j - \mu_k)(\xi_j - \xi_k)^2. \end{aligned} \tag{19}$$

Because of the assumption that $\psi''(r) > 0$ and the connectedness of the pixels, the second sum in (19) is negative unless $\xi_j = \xi_k$ for all j and k. If ξ_j is constant, substitution of this constant into the first sum of (19) shows that the first sum is negative.

For b) it suffices to prove that

$$\lim_{\|\mu\| \to \infty} L(\mu) \quad = \quad -\infty$$

since $U(\mu)$ is bounded below. Indeed, the bound $\inf_\mu U(\mu) > -\infty$ follows directly from the bound $\inf_r \psi(r) > -\infty$. The

limiting behavior of $L(\mu)$ holds because if any component μ_j tends to ∞, then the assumption $Y_i l_{ij} > 0$ for some projection i forces the conclusion $\lim_{n\to\infty} Y_i\langle l_i, \mu^n\rangle = \infty$.

Part c) follows from the fact that an unconstrained, strictly concave function can have at most one stationary point. Corresponding to each set of possible boundary restrictions $\mu_j = 0$, there is consequently at most one stationary point. ■

The next lemma states some properties of the iteration scheme μ^n.

Lemma 3 *Suppose that the iterations begin with μ^0 having all components positive. Then*

a) All components of each iterate μ^n are positive.

b) $\Delta(\mu^{n+1}) \geq \Delta(\mu^n)$, with strict inequality when $\mu^{n+1} \neq \mu^n$.

c) The iterates μ^n all belong to the same compact, convex set.

d) $\lim_{n\to\infty} \Delta(\mu^n)$ exists and is finite.

e) The Euclidean distance $\|\mu^{n+1} - \mu^n\|$ between successive iterates μ^{n+1} and μ^n tends to 0.

f) If some subsequence μ^{n_k} converges to μ^∞, then the subsequence μ^{n_k+1} also converges to μ^∞.

PROOF: Part a) follows directly from the definition of μ^{n+1} and the positivity of the components of μ^0. Part b) restates inequality (15). Part c) is true since all iterates belong to the set

$$\{\mu : \Delta(\mu) \geq \Delta(\mu^0)\}.$$

This set is compact because of the coerciveness of $\Delta(\mu)$ established in b) of Lemma 2. It is convex because $\Delta(\mu)$ is concave. Part d) follows from b) and the boundedness of $\Delta(\mu)$ on the compact set $\{\mu : \Delta(\mu) \geq \Delta(\mu^0)\}$.

To prove e) we expand $\Upsilon(\mu \mid \mu^n) - \Upsilon(\mu^n \mid \mu^n)$ in a second order Taylor's expansion around μ^{n+1}. If $d\Upsilon(\mu \mid \mu^n)$ and $d^2\Upsilon(\mu \mid \mu^n)$ denote the first and second differentials of $\Upsilon(\mu \mid \mu^n)$ with respect to its left argument, then

$$\begin{aligned} & \Upsilon(\mu^{n+1} \mid \mu^n) - \Upsilon(\mu^n \mid \mu^n) \\ = \;& d\Upsilon(\mu \mid \mu^n)\mid_{\mu=\mu^{n+1}} (\mu^{n+1} - \mu^n) \\ & -\frac{1}{2}(\mu^{n+1} - \mu^n)^t d^2\Upsilon(\mu \mid \mu^n)\mid_{\mu=\overline{\mu}} (\mu^{n+1} - \mu^n), \end{aligned} \tag{20}$$

where $\overline{\mu}$ is some point on the line segment between μ^n and μ^{n+1}. The linear term in (20) is nonnegative. This follows because the direction $\mu^n - \mu^{n+1}$ from μ^{n+1} is a descent direction. The quadratic form in (20) is positive definite and bounded below by the contribution of the Gibbs prior

$$\begin{aligned} & \sum_j \gamma \sum_{\{k:\{j,k\}\in N\}} w_{jk}\psi''(2\overline{\mu}_j - \mu_j^n - \mu_k^n)(\mu_j^{n+1} - \mu_j^n)^2 \\ \geq \;& c\|\mu^{n+1} - \mu^n\|^2. \end{aligned}$$

The constant c appearing in this last inequality is positive owing to part c) and the assumption that $\psi(r)$ is twice continuously differentiable and satisfies $\psi''(r) > 0$ for all r. Combining these developments with inequality (13) yields

$$\begin{aligned} \Delta(\mu^{n+1}) - \Delta(\mu^n) &\geq \Upsilon(\mu^{n+1} \mid \mu^n) - \Upsilon(\mu^n \mid \mu^n) \\ &\geq c\|\mu^{n+1} - \mu^n\|^2. \end{aligned}$$

Appeal to d) of the current lemma now finishes the proof of e). Part f) is an immediate consequence of e). ■

The preceding two lemmas set the stage for our proof of global convergence.

Theorem 1 *If the initial iterate μ^0 has all components positive, then the sequence μ^n converges to the global maximum of the log posterior $\Delta(\mu)$.*

PROOF: Because the sequence μ^n is confined to a compact set, it suffices to show that limit set of the sequence reduces to a single point and that this point is the maximum point. Suppose that $\mu^\infty = \lim_{k\to\infty} \mu^{n_k}$ is the limit of some subsequence μ^{n_k}. Let us first show that μ^∞ is a stationary point of $\Delta(\mu)$. As noted in c) of Lemma 2, we must demonstrate that all components of μ^∞ satisfy either $\mu_j^\infty = 0$ or $\frac{\partial}{\partial\mu_j}\Delta(\mu^\infty) = 0$. In the nontrivial case $\mu_j^\infty > 0$, the condition $\mu_j^{n_k+1} = \epsilon\mu_j^{n_k}$ cannot hold for infinitely many k since this would drive $\mu_j^{n_k+1}$ to 0 rather than to μ_j^∞, in contradiction to f) of Lemma 3. Thus $\mu_j^{n_k+1} = \hat{\mu}_j^{n_k}$ is true for all large k. It is then clear that the two equations

$$\begin{aligned} 0 &= \frac{\partial}{\partial\mu_j}\Upsilon(\mu \mid \mu^{n_k})\mid_{\mu=\hat{\mu}^{n_k}} \\ \frac{\partial}{\partial\mu_j}\Delta(\mu^{n_k}) &= \frac{\partial}{\partial\mu_j}\Upsilon(\mu \mid \mu^{n_k})\mid_{\mu=\mu^{n_k}} \end{aligned}$$

yield in the limit the desired condition $\frac{\partial}{\partial\mu_j}\Delta(\mu^\infty) = 0$.

Next observe that the limit set of μ^n is connected because of assertion e) of Lemma 3 [21]. Since the limit set is contained in the set of stationary points of $\Delta(\mu)$, and the stationary points are finite in number, connectedness demands that the limit set consist of a single stationary point.

Thus we may assume that $\lim_{n\to\infty}\mu^n = \mu^\infty$ exists. To prove that μ^∞ is the maximum point, it suffices to verify that each component μ_j^∞ satisfying $\mu_j^\infty = 0$ also satisfies the Kuhn-Tucker condition $\frac{\partial}{\partial\mu_j}\Delta(\mu^\infty) \leq 0$ [11]. If the contrary condition $\frac{\partial}{\partial\mu_j}\Delta(\mu^\infty) > 0$ holds for such a boundary component, then

$$\frac{\partial}{\partial\mu_j}\Upsilon(\mu \mid \mu^n)\mid_{\mu=\mu^n} = \frac{\partial}{\partial\mu_j}\Delta(\mu^n) \quad > \quad 0$$

holds for all large n. However, this situation entails

$$\begin{aligned} \mu_j^{n+1} &= \hat{\mu}_j^n \\ &\geq \mu_j^n, \end{aligned}$$

which clearly is in conflict with $\lim_{n\to\infty}\mu_j^n = 0$. This contradiction establishes that μ^∞ is the maximum point. ■

V. PERFORMANCE ON SIMULATIONS

In this section we describe some representative simulations demonstrating the relative convergence rates of the three algorithms. For these examples, we used the penalized versions (16)

and (18) of the algorithms with a simple quadratic smoothing prior of the form

$$U(\mu) \quad = \quad \gamma \sum_j \sum_{k \in \mathcal{N}_j} w_{jk} (\mu_j - \mu_k)^2,$$

where $\mathcal{N}_j$ denotes the usual 8 pixel neighborhood of the jth square pixel. Conventionally one sets the weights w_{jk} to 1 for horizontal and vertical neighbors and to $\frac{1}{\sqrt{2}}$ for diagonal neighbors. This choice leads to spatially-variant image resolution, so we used the modified weights described in [4] to make the resolution approximately uniform. We selected the regularization parameter γ as suggested by Fessler [4] to achieve a resolution of 2.5 pixels or 1.125cm full-width at half maximum (FWHM).

For testing the algorithms, we used the synthetic attenuation map shown in Figure 1, representing a human thorax with linear attenuation coefficients 0.0165/mm, 0.0096/mm, and 0.0025/mm for bone, soft tissue, and lungs, respectively. The image was decomposed into a 128 by 64 array of 4.5mm pixels. We simulated a PET transmission scan with 192 radial bins and 256 angles uniformly spaced over 180°. The l_{ij} factors correspond to 6mm wide strip integrals on 3mm center-to-center spacing. (This is an approximation to the ideal line integral that accounts for finite detector width.) The d_i factors were generated using pseudo-random log-normal variates with a standard deviation of 0.3, to account for detector efficiency variations, and scaled so that $\sum_i d_i \exp(-\langle l_i, \mu \rangle)$ was one million counts. Pseudo-random Poisson transmission projections Y_i were generated with means $d_i \exp(-\langle l_i, \mu \rangle)$.

We initialized the iterative algorithms with two different starting conditions μ^0. In the first case we started from the filtered backprojection (FBP) image shown in Figure 1, except that we first reset all attenuation values to no less than 0.01 of the maximum estimated value. We reconstructed the FBP image with a second order Butterworth filter at a resolution of 2.5 pixels or 1.125cm FWHM. In the second case we started from a uniform image with attenuation coefficient 0.008/mm. For the M step of the EM algorithm we employed Newton's method for each parameter [19]. For the gradient algorithm, we enforced monotonicity by repeatedly halving the step size until the objective function increased.

Figure 2 shows a plot of the increase in the log-posterior $\Delta(\mu^n) - \Delta(\mu^0)$ function versus iteration n for the algorithms initialized with the FBP image. The computation time per iteration varies among the algorithms, so a more objective comparison is total computation time, which is shown in Figure 3 as cumulative CPU time as measured on a DEC 3000/800 workstation. The EM algorithm requires N_0 exponentiations per iteration, where N_0 is the number of non-zero l_{ij} factors ($N_0 \approx 5.6 \times 10^6$ in this case). The convex algorithm and gradient algorithm only require N_p exponentiations per iteration, where N_p is the number of projections ($N_p = 256 \times 192 \approx 5 \times 10^4$ in this case). Thus, when measured against CPU time, the gradient algorithm and the convex algorithm approach the asymptote of the log-posterior much faster than the EM algorithm (although not 100 times faster since the inner products use much of the time per iteration).

Figures 4 and 5 are analogous to Figures 2 and 3 except that here the algorithms are initialized with the uniform image. In this case the convex algorithm converges faster than the gradient algorithm and the EM algorithm in terms of both CPU time and number of iterations.

It is obvious from these figures that conclusions drawn about convergence rates depend strongly on the starting conditions of the algorithms. In the above cases, the initial FBP image has much higher log-posterior than the initial uniform image, and subsequent iterations make smaller changes in the log-posterior. Thus the plots for starting with the FBP image are more related to asymptotic convergence rate, whereas the plots for starting with the uniform image measure initial performance of the algorithms far from the optimal point.

Maximizing the log-posterior is a surrogate for the real goal of producing better images. Figure 6 compares the images produced by the three algorithms after 15, 30, 60 and 110 seconds of CPU time. The images from the EM algorithm are blurry, reflecting slow convergence starting from an initial uniform image. The gradient and convex algorithms produce very similar images and obviously converge much faster than the EM algorithm. As can seen from a comparison of Figures 1 and 6, the maximum a posteriori images have fewer streak artifacts than the FBP image.

VI. Discussion

Because the EM algorithm for transmission tomography is beginning to see practical application [18, 19, 23], it is timely to review and compare its performance with competing algorithms. Our limited experience confirms the widespread impression that incorporating a smoothing prior enhances overall image quality. This practical improvement is consistent with the better theoretical behavior of the smoothed algorithms. For instance, sufficient smoothing automatically turns an ill-conditioned maximum likelihood problem into an well-conditioned maximum a posteriori problem.

The smoothed versions of the gradient algorithm (4) and the convex algorithm (7) appear to be considerably more efficient than the EM algorithm. This is not surprising in view of the larger number of exponentiations entailed by the EM algorithm. We anticipate that this superiority will continue to hold in other simulation trials. Because we understand its convergence behavior better, we tend to prefer the convex algorithm to the gradient algorithm.

The convex and gradient algorithms should adapt well to array and parallel processing. A substantial proportion of the computation load for both algorithms involves calculation of the discrete line integrals $\langle l_i, \mu \rangle$. These and subsequent operations are perfect candidates for array and parallel processing. The EM algorithm, in contrast, is more awkward to implement since it involves sequential calculation of many partial line integrals. Of course, the algorithm of choice depends on the intended computer. It is noteworthy that on conventional serial workstations a non-parallelizable coordinate ascent algorithm converges faster from a FBP starting image than any the three algorithms examined here [5]

The EM and convex algorithms, and possibly the gradient

algorithm as well, could benefit from the quasi-Newton acceleration techniques recently suggested by Lange [15]. These techniques attempt to build better approximations to the Hessian $d^2\Delta$ of the log posterior using the diagonal Hessian $d^2\Upsilon$ of the comparison function Υ as a base. The presence of boundary constraints on the parameters complicates quasi-Newton methods, but perhaps the addition of small barrier terms to the log-likelihood will make acceleration techniques practical without detracting much from the final image.

Although the algorithms discussed here show definite promise, further theoretical improvements are to be expected. At the same time computing costs continue to drop, and processor speeds to increase. These trends imply an accelerating transition away from Fourier methods and toward statistical methods of image reconstruction.

REFERENCES

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

[2] A. R. De Pierro, "On the relation between the ISRA and the EM algorithm for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. 12, pp. 328-333, 1993.

[3] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imag.*, 1995 (in press).

[4] J. A. Fessler and W. L. Rogers, "Uniform quadratic penalties cause nonuniform image resolution (and sometimes vice versa)," *1994 IEEE Nuclear Symposium and Medical Imaging Conference*, 1994.

[5] J. A. Fessler, "Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans," *IEEE Trans. Im. Proc.*, Oct. 1995. To appear.

[6] S. Geman and D. McClure, "Bayesian image analysis: An application to single photon emission tomography," *Proc. Stat. Comput. Sec.*, American Statistical Association, Washington, DC, 12–18, 1985.

[7] S. Geman and D. McClure, "Statistical methods for tomographic image reconstruction," ISI Tokyo session. *Bull. Int. Stat. Inst.*, vol. 52, pp. 5–21, 1987.

[8] P. Green, "Bayesian reconstruction for emission tomography data using a modified EM algorithm," *IEEE Trans. Med. Imag.* vol. 9, pp. 84–94, 1990.

[9] P. J. Green, "On the use of the EM algorithm for penalized likelihood estimation," *J. Royal Stat. Soc. Series B*, vol. 52, pp. 443–452, 1990.

[10] G. T. Herman, *Image Reconstruction from Projections: the Fundamentals of Computerized Tomography*. Springer, New York, 1980.

[11] M. R. Hestenes, *Optimization Theory: The Finite Dimensional Case*. Robert E. Krieger Publishing Co., Huntington, New York, p. 186, 230–234, 1981.

[12] K. Lange, "An overview of Bayesian methods in image reconstruction," *Digital Image Synthesis and Inverse Optics*, edited by A. F. Gmitro, P. S. Idell, I. S. La Haie, Society of Photo-Optical Engineering, Bellingham, Washington, SPIE vol. 1351, pp. 270–287, 1990.

[13] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, pp. 439–446, corrections ibid. vol. 10, p. 228, 1990.

[14] K. Lange, "A gradient algorithm locally equivalent to the EM algorithm," *J. Royal Stat. Soc. Series B*, 1995 (in press).

[15] K. Lange, "A quasi-Newton acceleration of the EM algorithm," *Statistica Sinica*, vol. 5, 1995 (in press).

[16] K. Lange, M. Bahn, and R. Little, "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Trans. Med. Imag.*, vol. 6, pp. 106–114, 1987.

[17] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assist. Tomog.*, vol. 8, pp. 306–316, 1984.

[18] S. H. Manglos, F. D. Thomas, G. M. Gagne, and B. J. Hellwig, "Phantom study of breast tissue attenuation in myocardial imaging," *J. Nuclear Med.*, vol. 34, pp. 992–996, 1993.

[19] J. M. Ollinger, "Maximum likelihood reconstruction of transmission images in emission computed tomography via the EM algorithm," *IEEE Trans. Med. Imag.*, vol. 13, pp. 89–101, 1994.

[20] J. M. Ortega, *Numerical Analysis: A Second Course*. SIAM, Philadelphia, pp. 144–145, 1990.

[21] A. M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*, Academic, New York, p. 173, 1973.

[22] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. 1, pp. 113–121, 1982.

[23] C. H. Tung, G. T. Gullberg, G. L. Zeng, P. E. Christian, F. L. Datz, and H. T. Morgan, "Nonuniform attenuation correction using simultaneous transmission and emission converging tomography," *IEEE Trans. Nuclear Science*, vol. 39, pp. 1134–1143, 1992.

[24] Y. Vardi, L. A. Shepp, and L. Kaufman, "A statistical model for positron emission tomography," *J. Amer. Stat. Assoc.*, vol. 80, pp. 8–37, 1985.

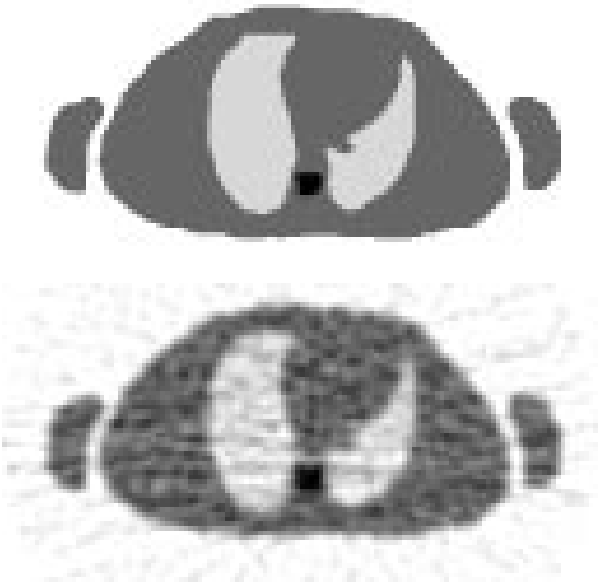

Figure 1: Digital thorax phantom (top), and image reconstructed using filtered backprojection (bottom).

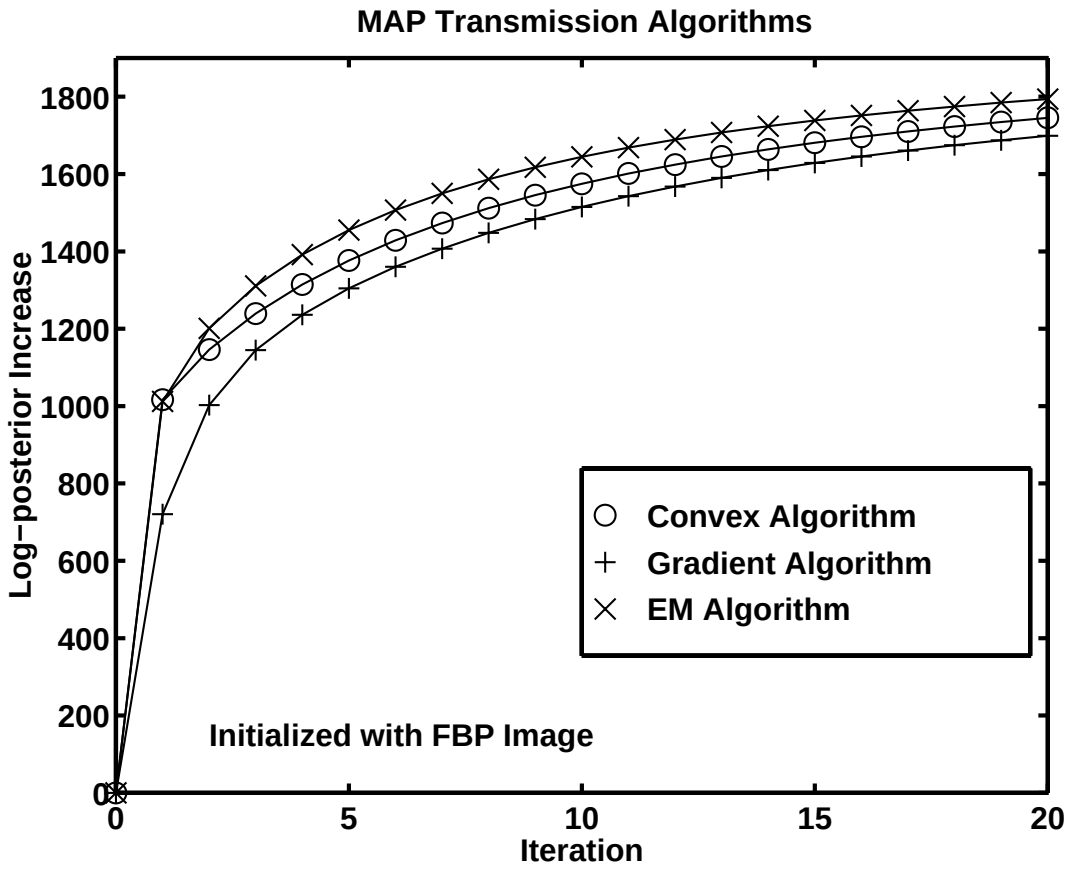

Figure 2: Increase in log-posterior $\Delta(\mu^n) - \Delta(\mu^0)$ versus iteration starting with the FBP image.

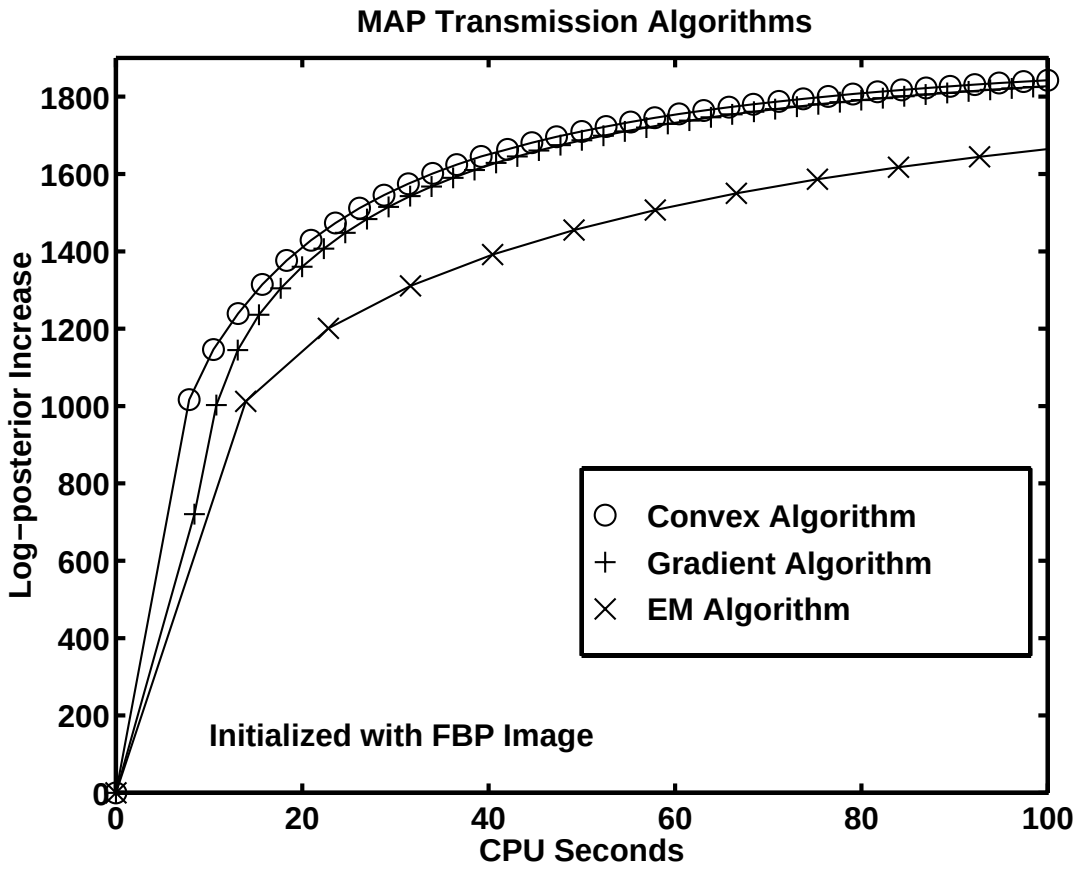

Figure 3: Increase in log-posterior $\Delta(\mu^n) - \Delta(\mu^0)$ versus CPU seconds starting with the FBP image.

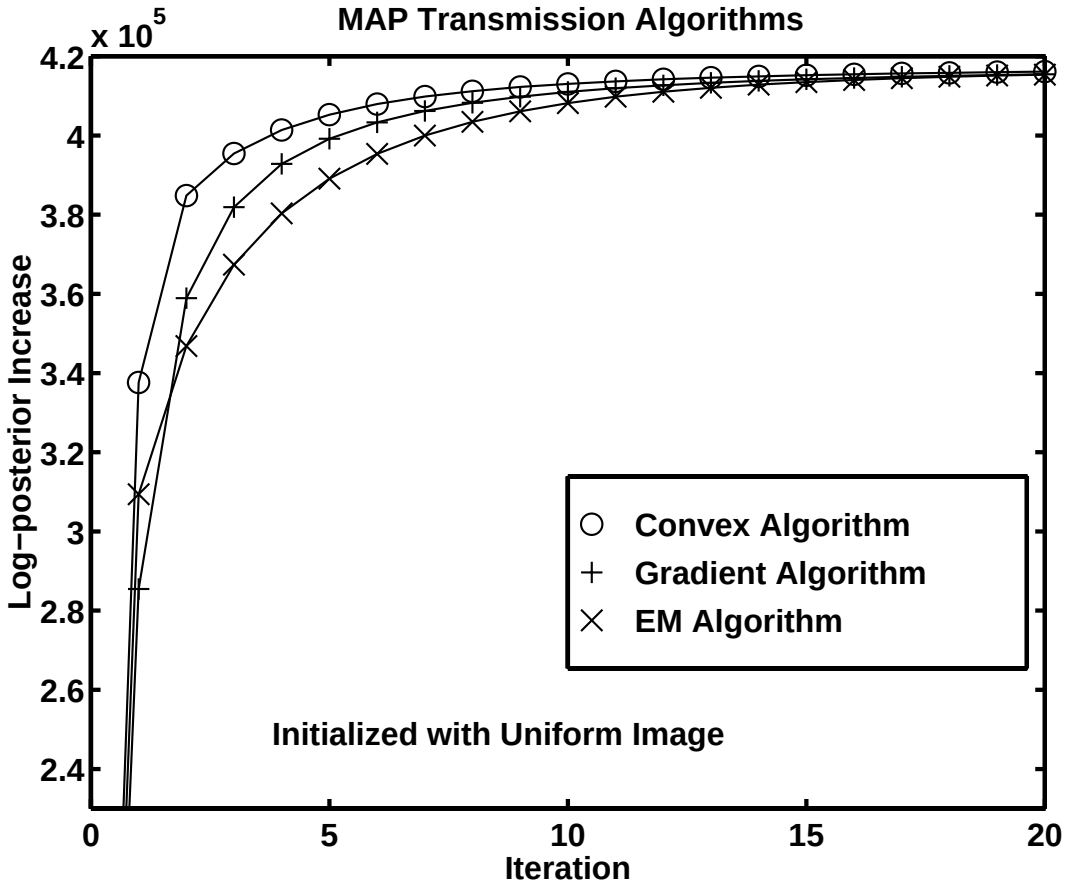

Figure 4: Increase in log-posterior $\Delta(\mu^n) - \Delta(\mu^0)$ versus iteration starting with the uniform image.

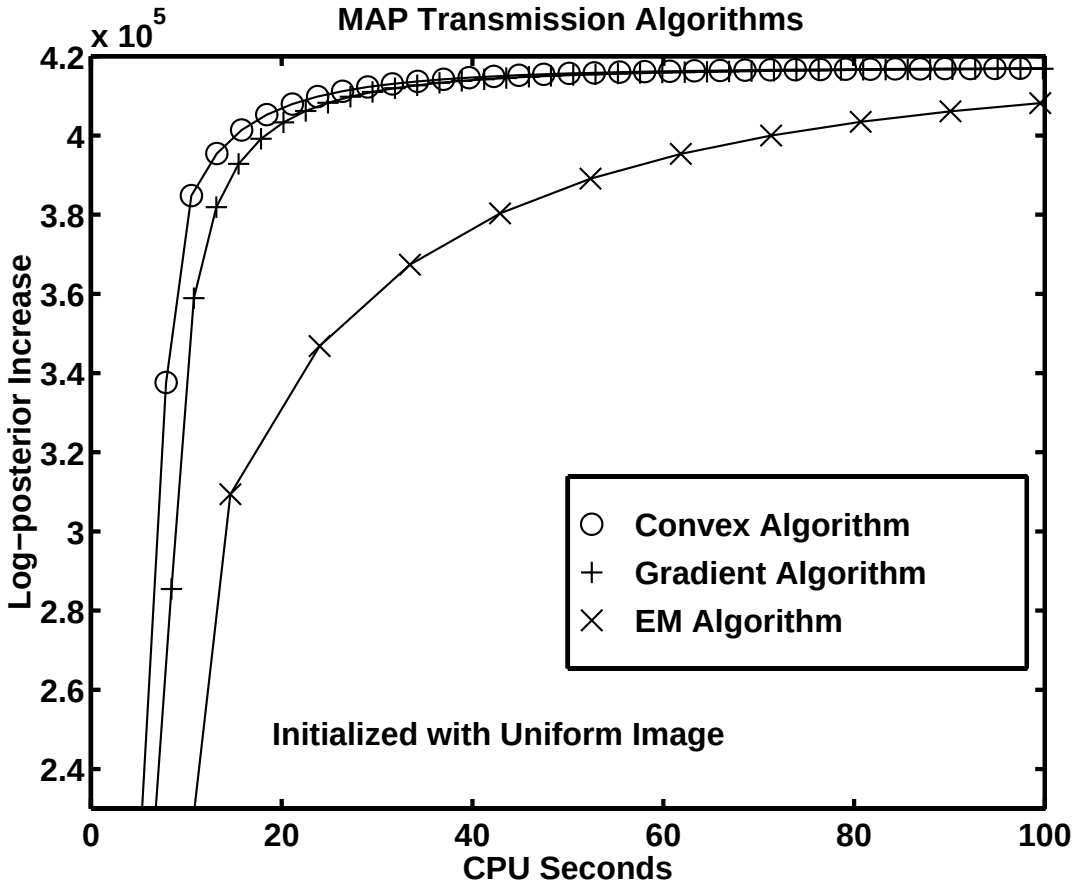

Figure 5: Increase in log-posterior $\Delta(\mu^n) - \Delta(\mu^0)$ versus CPU seconds starting with the uniform image.

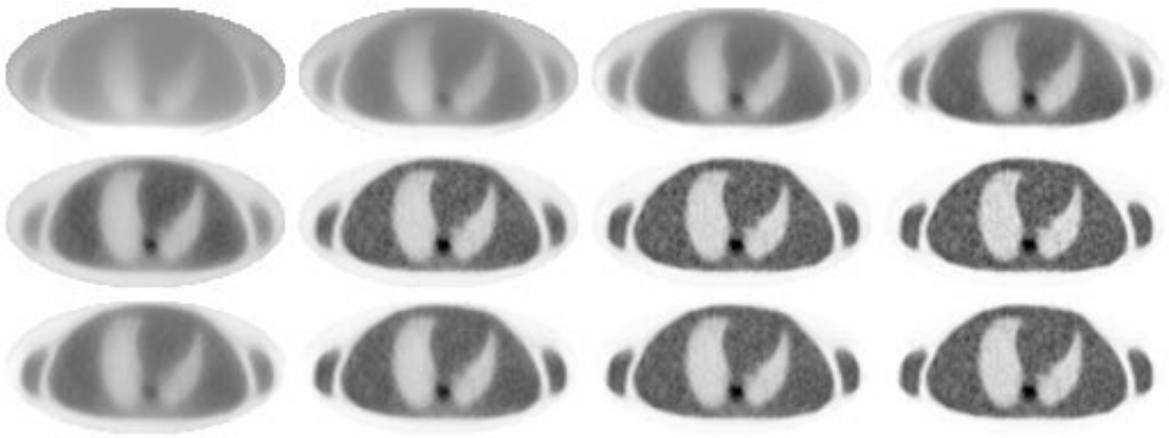

Figure 6: Images reconstructed after approximately 15, 30, 60, and 110 CPU seconds (left to right) when initialized with the uniform image. Top: iterations 1, 3, 6, and 11 of the EM algorithm. Middle: iterations 4, 12, 24, and 44 of the gradient algorithm. Bottom: iterations 4, 12, 22, and 40 of the convex algorithm.