

Grouped-Coordinate Ascent Algorithms for Penalized-Likelihood Transmission Image Reconstruction

Jeffrey A. Fessler, Edward P. Ficaro, Neal H. Clinthorne, and Kenneth Lange
4240 EECS Bldg., University of Michigan, Ann Arbor, MI 48109-2122
Email: fessler@umich.edu, Voice: 734-763-1434

IEEE T-MI 16(2):166-75, Apr. 1997

ABSTRACT

This paper presents a new class of algorithms for penalized-likelihood reconstruction of attenuation maps from low-count transmission scans. We derive the algorithms by applying to the transmission log-likelihood a version of the convexity technique developed by De Pierro for emission tomography. The new class includes the single-coordinate ascent (SCA) algorithm and Lange's convex algorithm for transmission tomography as special cases. The new grouped-coordinate ascent (GCA) algorithms in the class overcome several limitations associated with previous algorithms. (1) Fewer exponentiations are required than in the transmission ML-EM algorithm or in the SCA algorithm. (2) The algorithms intrinsically accommodate nonnegativity constraints, unlike many gradient-based methods. (3) The algorithms are easily parallelizable, unlike the SCA algorithm and perhaps line-search algorithms. We show that the GCA algorithms converge faster than the SCA algorithm, even on conventional workstations. An example from a low-count positron emission tomography (PET) transmission scan illustrates the method.

I. INTRODUCTION

STATISTICAL methods for reconstructing attenuation images from transmission scans have increased in importance recently for several reasons, including the necessity of reconstructing 2D attenuation maps for reprojection to form 3D attenuation correction factors in septaless PET [1, 2], the widening availability of SPECT systems equipped with transmission sources [3], and the potential for reducing transmission noise in whole body PET images and in other protocols requiring short transmission scans [4]. The non-statistical filtered back-projection (FBP) method and the data-weighted least-squares method [5] for transmission image reconstruction lead to systematic biases for low-count scans [6–8]. These biases are due to the nonlinearity of the logarithm applied to the transmission data. To eliminate these biases, one can use statistical methods based on the Poisson measurement statistics, which use the raw measurements rather than its logarithms [6, 9–11]. Statistical methods also offer reduced variance relative to FBP [6, 8, 12].

Several reconstruction algorithms based on the Poisson statistical model for transmission scans [13] have appeared recently [6, 10, 11, 14–20], all of which converge faster than the original transmission maximum-likelihood expectation-maximization (ML-EM) algorithm [9]. Nevertheless, each of these methods is still less than ideal due to one or more of the following reasons.

- The EM algorithms [9, 18] and single-coordinate ascent (SCA) algorithms [6, 10, 11] require at least one exponentiation per nonzero element of the system matrix per iteration, which is a large computational expense.
- Enforcing nonnegativity in gradient-based algorithms [19–22] is possible but somewhat awkward.
- Many algorithms are poorly suited to parallel processors such as the i860 arrays that are common at septaless PET sites. This is true of SCA methods and of algorithms that use line searches, since a line-search step may not parallelize easily.

This paper describes a new class of algorithms for reconstructing attenuation maps from low-count transmission scans. These algorithms are parallelizable, easily accommodate nonnegativity constraints and nonquadratic convex penalties, and require a moderate number of exponentiations. The derivation of these transmission algorithms exploits two ideas underlying recent developments in algorithms for emission tomography: updating the parameters in groups [23, 24], and the convexity technique of De Pierro [25, 26]. Integrating these two ideas leads to a new class of algorithms [27] that converge quickly and with less computation than previous statistical methods for transmission tomography.

This work can be considered a generalization of previous methods for tomographic image reconstruction based on *sequential* updates [5, 10, 11, 23, 24, 28, 29]. The fast convergence of sequential updates for tomographic problems was analyzed by Fourier methods and shown empirically to converge faster than *simultaneous* updates in [5]. Tomographic reconstruction is an important case of the general problem of estimating superimposed signals [30–33]. In [31] a sequential method called the “alternating maximization” (AM) algorithm was proposed for this estimation problem, whereas [32] proposed a simultaneous update based on an EM algorithm. The improved asymp-

This work was supported in part by NIH grants CA-60711 and CA-54362. Permission granted to publish this abstract separately.

otic convergence rate of the sequential AM algorithm relative to the simultaneous EM algorithm was shown in Chapter 5 of [34] (under somewhat restrictive conditions) and later generalized in [23]. Such sequential algorithms have been given many names, including iterated conditional modes [35], Gauss-Siedel [5, 28], successive over-relaxation [29], cyclic coordinate ascent [6], and iterative coordinate descent [11, 36]. In this paper we use the names *single-coordinate ascent* and *grouped-coordinate ascent* to distinguish the case where one pixel at a time is updated from the parallelizable case where several pixels are updated simultaneously.

After submitting the abstract for [27], we learned of the independent work of Sauer *et al.* [37], which includes an algorithm that is similar to one of the algorithms in the class proposed here. The emphasis in [37] is on the parallelizability of the algorithms. In this paper, we emphasize the point that, when implemented efficiently, the new class of algorithms leads to faster computation *even on a conventional single-processor workstation*. This paper also considers random coincidences, unlike [27, 37]. Finally, unlike in [37], we do not make a one-time quadratic approximation to the log-likelihood, since that approximation can lead to systematic biases for low-count PET and SPECT transmission scans [6, 8].

There has also been work on grouped-coordinate ascent algorithms in the statistics literature [38], which in turn cites related algorithms dating to 1964! So clearly what is new in this paper is not the general idea of updating parameters sequentially or in groups, but rather is the specifics of how the iterations and updates can be formulated to achieve a reasonable balance between *convergence rate* and *computation per iteration* in the PET transmission problem.

The remainder of this paper describes the problem, develops the new algorithms, and presents a representative example of performance on real PET transmission data.

II. PROBLEM

The Poisson statistical model is widely used for transmission measurements that have been formed by counting individual photons (SPECT) or photon pairs (PET). In practice, both SPECT and PET transmission measurements also contain extra counts due to “background” events such as random coincidences [39], scatter [40], emission crosstalk [3], and room background. We assume

$$y_i \sim \text{Poisson}\{b_i e^{-\langle a_i, \theta_{\text{true}} \rangle} + r_i\}, \quad i = 1, \dots, N, \quad (1)$$

where N is the number of measurements,

$$\langle a_i, \theta \rangle = \sum_{j=1}^p a_{ij} \theta_j$$

represents the i th “line integral” through the attenuation map, a_i denotes the i th row of the $N \times p$ system matrix $\mathbf{A} = \{a_{ij}\}$, y_i denotes the transmission measurement of the i th detector, b_i denotes the i th blank scan measurement, r_i denotes the mean number of background counts in the i th measurement, θ_j denotes the unknown attenuation coefficient in the j th voxel (units: inverse

length), and the a_{ij} ’s are the transmission system model (units: length) [41]. We assume $\{b_i\}$, $\{r_i\}$, and $\{a_{ij}\}$ are known non-negative constants.

For independent transmission measurements, the log-likelihood is [9]:

$$L(\theta) = \sum_{i=1}^N h_i(\langle a_i, \theta \rangle) \quad (2)$$

where (neglecting constants independent of θ hereafter):

$$h_i(l) = y_i \log(b_i e^{-l} + r_i) - (b_i e^{-l} + r_i). \quad (3)$$

The algorithms developed below apply to any problem of the form (2) with concave h_i , including “data-weighted” least squares estimation [6], the “estimate-weighted” least squares objective function described in [42], and penalized-likelihood emission tomography [37].

Since maximizing $L(\cdot)$ leads to unacceptably noisy images, our goal is to compute a penalized-likelihood estimate $\hat{\theta}$ of the attenuation map θ^{true} , with $\hat{\theta}$ defined by

$$\hat{\theta} = \arg \max_{\theta \geq 0} \Phi(\theta), \quad \Phi(\theta) = L(\theta) - \beta R(\theta), \quad (4)$$

where the objective includes a roughness penalty

$$R(\theta) = \sum_j \frac{1}{2} \sum_k w_{jk} \psi(\theta_j - \theta_k). \quad (5)$$

The function ψ should be symmetric and twice differentiable. Ordinarily $w_{jk} = 1$ for horizontal and vertical neighboring pixels, $w_{jk} = 1/\sqrt{2}$ for diagonal neighboring pixels, and $w_{jk} = 0$ otherwise. For the results in Section V we adopt the modification described in [12, 43, 44], which provides more uniform spatial resolution.

A. Penalty Function

Although the method applies more generally, for concreteness in this paper we focus on one of the penalties proposed in [16]:

$$\psi(x) = \delta^2 [|x/\delta| - \log(1 + |x/\delta|)]. \quad (6)$$

This function approaches $\psi(x) = x^2/2$ as $\delta \rightarrow \infty$, but provides a degree of edge preservation for finite δ . Since

$$\dot{\psi}(x) = \frac{d}{dx} \psi(x) = \frac{x}{1 + |x/\delta|} \quad (7)$$

implies $|\dot{\psi}(x)| < \delta$, this potential function has bounded influence. The derivative (7) of ψ requires no transcendental functions, which is desirable computationally.

B. Concavity

When $r_i = 0$, each function h_i is concave over all of \mathbb{R} , so it is easily verified that $L(\cdot)$ is concave over all of \mathbb{R}^p . Since ψ is strictly convex and $L(\cdot)$ is concave, the objective Φ is strictly concave under mild conditions on \mathbf{A} [20]. This concavity is central to the development of the algorithms below.

C. Why another algorithm?

Direct maximization of (4) is intractable, so one must use iterative algorithms. Generic numerical methods such as steepest ascent do not exploit the specific structure of Φ , nor do they easily accommodate nonnegativity constraints. Thus for fastest convergence, one must seek algorithms tailored to this problem. Relevant properties of L include:

- $L(\theta)$ is a sum of concave functions $h_i(\cdot)$ (when $r_i = 0$).
- The arguments of the functions h_i are inner products.
- The inner product coefficients are all nonnegative.

These properties suggest the use of Jensen's inequality.

III. GROUPED-COORDINATE ASCENT ALGORITHMS

As shown by frequency domain analysis in [5], sequential updates such as SCA converge very rapidly for tomographic reconstruction. Unfortunately, for transmission tomography the SCA update requires a large number of exponentiations. Consider the partial derivative of the log-likelihood (2) with respect to the j th pixel value:

$$\dot{L}_j(\theta) = \frac{\partial}{\partial \theta_j} L(\theta) = \sum_{i \in \mathcal{I}_j} a_{ij} \left[1 - \frac{y_i}{\bar{y}_i(\theta)} \right] b_i e^{-\langle a_i, \theta \rangle}, \quad (8)$$

where

$$\bar{y}_i(\theta) = b_i e^{-\langle a_i, \theta \rangle} + r_i$$

(see Eqn. (8) of [6]), and where

$$\mathcal{I}_j = \{i : a_{ij} \neq 0\}.$$

A SCA algorithm must repeatedly evaluate $\dot{L}_j(\theta^n)$ at the current image estimate θ^n . Since $\langle a_i, \theta^n \rangle$ changes immediately after each pixel is updated, one can see from (8) that each complete iteration requires M exponentiations, where M is the number of nonzero a_{ij} 's. At the other extreme, Lange's convex algorithm [15, 20] and scaled-gradient algorithm [16, 20] for transmission tomography update all pixels simultaneously. Thus one can compute simultaneously the i -subscripted terms in (8) prior to the backprojection in (8), so only N exponentiations are required. Typically the number of measurements N is two orders of magnitude smaller than M . In other words, there is an "economy of scale" in terms of computation by updating all pixels simultaneously². However, simultaneous updates lead to slow convergence rates [5, 6, 23, 24].

Rather than updating *all* pixels simultaneously, we propose to update only certain *groups of pixels* simultaneously. If one uses G groups of pixels, then only NG exponentiations are needed (see Table 1). On the other hand, if the pixels in each group are well-separated spatially, then we anticipate that they will be fairly well decoupled, so the algorithm will not suffer from slow convergence. The results in Section V confirm this intuition.

²Even if the exponentiations are computed approximately, using table lookups for example, the ratio between N and M remains unchanged.

Let \mathcal{S} be a subset of the pixels $\{1, \dots, p\}$, let $\bar{\mathcal{S}}$ be its complement, and let $|\mathcal{S}|$ be the cardinality of \mathcal{S} . In a GCA algorithm³, we update $\theta_{\mathcal{S}}$ while holding $\theta_{\bar{\mathcal{S}}}^n$ fixed at the n th update [23]. Unfortunately it is even too difficult to maximize $\Phi(\theta_{\mathcal{S}}, \theta_{\bar{\mathcal{S}}}^n)$ over $\theta_{\mathcal{S}}$ directly, so we will settle for finding a method for choosing $\theta_{\mathcal{S}}^{n+1}$ that will at least *monotonically increase* the objective function⁴:

$$\Phi(\theta_{\mathcal{S}}^{n+1}, \theta_{\bar{\mathcal{S}}}^n) \geq \Phi(\theta_{\mathcal{S}}^n, \theta_{\bar{\mathcal{S}}}^n) = \Phi(\theta^n).$$

To ensure monotonicity, we use a generalization of De Pierro's *optimization transfer* idea [25, 26], which is illustrated in Fig. 1. Instead of trying to find $\theta_{\mathcal{S}}^{n+1}$ to maximize $\Phi(\theta_{\mathcal{S}}, \theta_{\bar{\mathcal{S}}}^n)$, we maximize a *surrogate function* $\phi(\theta_{\mathcal{S}}; \theta^n)$ over a corresponding *region of monotonicity* $\mathcal{R}_{\mathcal{S}}(\theta^n) \subseteq \mathbb{R}^{|\mathcal{S}|}$ that we must choose to satisfy:

$$\Phi(\theta_{\mathcal{S}}, \theta_{\bar{\mathcal{S}}}^n) - \Phi(\theta^n) \geq \phi(\theta_{\mathcal{S}}; \theta^n) - \phi(\theta_{\mathcal{S}}^n; \theta^n), \quad \forall \theta_{\mathcal{S}} \in \mathcal{R}_{\mathcal{S}}(\theta^n). \quad (9)$$

The GCA update (cf SAGE algorithm [23, 24]) is then:

$$\begin{aligned} \theta_{\mathcal{S}}^{n+1} &= \arg \max_{\theta_{\mathcal{S}} \in \mathcal{R}_{\mathcal{S}}(\theta^n)} \phi(\theta_{\mathcal{S}}; \theta^n), \\ \theta_j^{n+1} &= \theta_j^n, \quad j \in \bar{\mathcal{S}}. \end{aligned} \quad (10)$$

The condition (9) is sufficient to ensure that the iterates produced by the above generic update will monotonically increase the objective: $\Phi(\theta^{n+1}) \geq \Phi(\theta^n)$.

A. Choosing Surrogate Functions

We restrict attention here to additively separable⁵ surrogate functions $\phi(\cdot; \theta^n)$ satisfying

$$\phi(\theta_{\mathcal{S}}; \theta^n) = \sum_{j \in \mathcal{S}} \phi_j(\theta_j; \theta^n). \quad (11)$$

To choose these ϕ_j 's, we use modifications of De Pierro's convexity method [25, 26] rather than the EM approach of [23, 24]. The key step is to note that

$$\langle a_i, [\theta_{\mathcal{S}}, \theta_{\bar{\mathcal{S}}}^n] \rangle = \sum_{j \in \mathcal{S}} \alpha_{ij} \left[\frac{a_{ij}}{\alpha_{ij}} (\theta_j - \theta_j^n) + \langle a_i, \theta^n \rangle \right] \quad (12)$$

for any choice⁶ of $\alpha_{ij} \geq 0$ that satisfies the constraint

$$\sum_{j \in \mathcal{S}} \alpha_{ij} = 1, \quad \forall i. \quad (13)$$

We discuss specific choices for α_{ij} in the next section.

³In a GCA method, \mathcal{S} varies with n . To simplify notation, we leave this dependence implicit.

⁴To simplify notation, in the presentation we increment n every time a group of pixels is updated. We reserve the term "iteration" to mean a complete update of all pixels.

⁵Separable surrogate functions are very convenient for enforcing the non-negativity constraint. There may be alternatives that lead to faster convergence though.

⁶We assume $\alpha_{ij} = 0$ if and only if $a_{ij} = 0$ so that (12) is well defined.

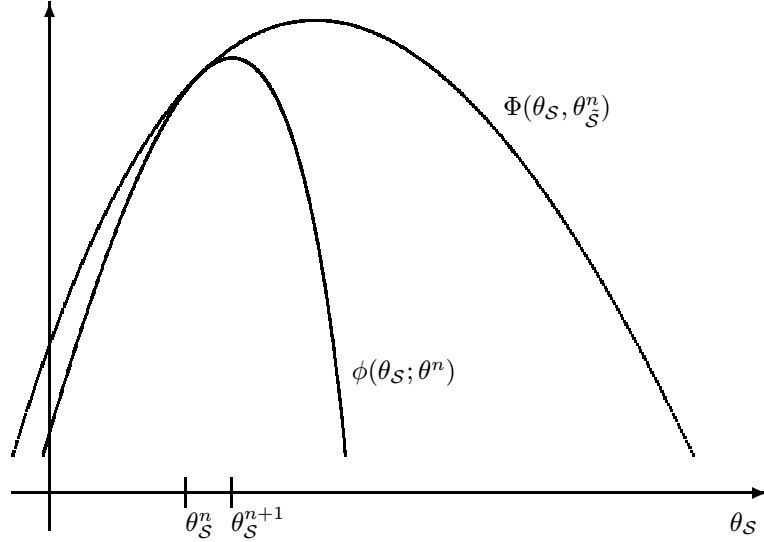


Figure 1: One-dimensional illustration of the optimization transfer principle: instead of maximizing $\Phi(\theta_S, \theta_S^n)$ over θ_S , we maximize the surrogate function $\phi(\theta_S, \theta^n)$ iteratively. The higher the curvature of $\phi(\cdot, \theta^n)$, i.e. the greater the norm of its Hessian, the slower the convergence rate [23, 45].

When h_i is concave over all of \mathbb{R} (such as when $r_i = 0$), then it follows directly from (12) and the convexity equality that:

$$h_i(\langle a_i, [\theta_S, \theta_S^n] \rangle) \geq \sum_{j \in \mathcal{S}} \alpha_{ij} h_i \left(\frac{\alpha_{ij}}{\alpha_{ij}} (\theta_j - \theta_j^n) + \langle a_i, \theta^n \rangle \right). \quad (14)$$

Unfortunately, when r_i is nonzero, h_i is concave only over the interval $(-\infty, l_i^{\max})$ where (see [6] or (22) below):

$$l_i^{\max} = \begin{cases} \infty, & r_i = 0 \text{ or } r_i \geq y_i \\ \log \left(\frac{b_i}{\sqrt{y_i r_i} - r_i} \right), & \text{otherwise} \end{cases}.$$

Thus the inequality in (14) is guaranteed (by the convexity inequality) to be satisfied only for θ_S such that

$$\frac{\alpha_{ij}}{\alpha_{ij}} (\theta_j - \theta_j^n) + \langle a_i, \theta^n \rangle \leq l_i^{\max}, \quad \forall j \in \mathcal{S}, \forall i \in \mathcal{I}_j.$$

Consequently, we define $\mathcal{R}_S(\theta^n)$ as follows:

$$\mathcal{R}_S(\theta^n) = \{\theta_S \geq 0 : \theta_j \leq \theta_{j, \max}^n, \forall j \in \mathcal{S}\}, \quad (15)$$

where

$$\theta_{j, \max}^n = \arg \min_{i \in \mathcal{I}_j} \left\{ \theta_j^n + \frac{\alpha_{ij}}{\alpha_{ij}} (l_i^{\max} - \langle a_i, \theta^n \rangle) \right\}. \quad (16)$$

For typical small values of r_i , it is reasonable to expect that $l_i^{\max} \gg \langle a_i, \theta^n \rangle$, so $\mathcal{R}_S(\theta^n)$ will contain most of the relevant part of $\mathbb{R}^{|\mathcal{S}|}$.

Using the definition (15) as our region of monotonicity $\mathcal{R}_S(\theta^n)$, it follows from (14) that we have:

$$L([\theta_S, \theta_S^n]) = \sum_{i=1}^N h_i(\langle a_i, [\theta_S, \theta_S^n] \rangle) \geq \sum_{j \in \mathcal{S}} Q_j(\theta_j; \theta^n)$$

for $\theta_S \in \mathcal{R}_S(\theta^n)$, where using (12) and (14):

$$Q_j(\theta_j; \theta^n) = \sum_{i \in \mathcal{I}_j} \alpha_{ij} h_i \left(\frac{\alpha_{ij}}{\alpha_{ij}} (\theta_j - \theta_j^n) + \langle a_i, \theta^n \rangle \right). \quad (17)$$

Assuming the groups are chosen so that no two neighboring pixels are in the same group⁷, then the surrogate function defined by (11) with⁸

$$\phi_j(\theta_j; \theta^n) = Q_j(\theta_j; \theta^n) - \beta \sum_k w_{jk} \psi(\theta_j - \theta_k^n) \quad (18)$$

will satisfy the monotonicity condition (9). Each ϕ_j only depends on one θ_j , so since $\mathcal{R}_S(\theta^n)$ defined in (15) above is separable, the maximization step in (10) reduces to $|\mathcal{S}|$ separate 1D maximizations. Thus (10) becomes the parallelizable operations:

$$\theta_j^{n+1} = \arg \max_{0 \leq \theta_j \leq \theta_{j, \max}^n} \phi_j(\theta_j; \theta^n), \quad j \in \mathcal{S}. \quad (19)$$

B. Convergence

When $r_i = 0 \forall i$ so that Φ is globally strictly concave, it is fairly straightforward to apply the general convergence proof in [23] to prove that the sequence of estimates $\{\theta^n\}$ produced by the above algorithm ((10) and (19)) monotonically increases Φ and converges from any starting image to the unique global maximizer of Φ subject to $\theta \geq 0$, under mild assumptions about \mathcal{S} and the α_{ij} 's. There are a few practical caveats that should

⁷If a group contains neighboring pixels, then one can also apply De Pierro's penalty function approach [25, 26] to ensure (9). For a first or second order neighborhood, the only change is a factor of 2 following the parameter β in (26) and in the denominator of (29) [6].

⁸Note that the $\frac{1}{2}$ in (5) disappears in (18) since each pair of pixels is counted twice in (5).

be considered however. When using finite precision arithmetic, monotonicity may not hold exactly when the sequence gets very close to the maximum. Also, usually one will not perform exact 1D maximizations as implied by (19), but rather partial or approximate maximizations (see below). Finally, when $r_i \neq 0$, it is cumbersome to compute the $\theta_{j,\max}^n$ terms, so in our software we take the more pragmatic approach of simply verifying that Φ has increased after each complete iteration. (We have yet to observe nonmonotonicity exceeding numerical precision limits in thousands of reconstructions.) Verifying monotonicity does not ensure global convergence in the non-concave case. Nevertheless, it is comforting to know that, at least under ideal circumstances (i.e. $r_i = 0$, perfect numerical precision, exact maximizations), global convergence is ensured.

C. The Maximization Step

One simple approach to implementing the maximization (19) would be to apply a few sub-iterations of the 1D Newton-Raphson method:

$$\begin{aligned} \theta_j^{\text{work}} &= \theta_j^n \\ \theta_j^{\text{work}} &:= \left[\theta_j^{\text{work}} + \frac{\frac{d}{d\theta_j} \phi_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^{\text{work}}}}{-\frac{d^2}{d\theta_j^2} \phi_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^{\text{work}}}} \right]_+ \\ \theta_j^{n+1} &= \theta_j^{\text{work}}, \end{aligned} \quad (20)$$

where $[x]_+ = x$ for $x > 0$ and is 0 otherwise. This $[\cdot]_+$ operator enforces the nonnegativity constraint. The “:=” symbol in the middle step above indicates “in place” computation, and typically this step would be repeated a few times. Unfortunately, the partial derivatives of $\phi_j(\cdot; \theta^n)$ are fairly expensive to compute exactly, so (20) is impractical.

To reduce computation, we apply methods from [11] and [6]. For the numerator, we approximate the Q_j function in (17) (but not the penalty!) by its second-order Taylor series about the current estimate θ_j^n , in a spirit similar to [11]. For the denominator, we use a trick similar to [6] for precomputing an approximation to the second derivative of the Q_j function, and a new trick for the penalty term that exploits its bounded curvature.

The second-order Taylor expansion about θ_j^n for the $Q_j(\cdot; \theta^n)$ component of the numerator is:

$$Q_j(\theta_j; \theta^n) \approx Q_j(\theta_j^n; \theta^n) + \dot{L}_j(\theta^n)(\theta_j - \theta_j^n) - \frac{d_j(\theta^n)}{2}(\theta_j - \theta_j^n)^2,$$

because from (17) it follows that

$$\frac{d}{d\theta_j} Q_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^n} = \frac{\partial}{\partial \theta_j} L(\theta) \Big|_{\theta = \theta^n} = \dot{L}_j(\theta^n),$$

and that

$$d_j(\theta^n) = - \frac{d^2}{d\theta_j^2} Q_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^n} = - \sum_{i \in \mathcal{I}_j} \frac{a_{ij}^2}{\alpha_{ij}} \ddot{h}_i(\langle a_i, \theta^n \rangle), \quad (21)$$

where (see [6]):

$$-\ddot{h}_i(l) = \left[1 - \frac{y_i r_i}{(b_i e^{-l} + r_i)^2} \right] b_i e^{-l}. \quad (22)$$

Note that θ^n only enters $d_j(\theta^n)$ through its projections $\langle a_i, \theta^n \rangle$. Thus $d_j(\theta^n)$ is fairly insensitive to θ^n , so we replace $\langle a_i, \theta^n \rangle$ with a precomputed approximation to the i th line integral, such as $\log(b_i/(y_i - r_i))$. Specifically we replace $d_j(\theta^n)$ with the approximation:

$$\hat{d}_j = - \sum_{\substack{i \in \mathcal{I}_j \\ y_i \neq 0}} \frac{a_{ij}^2}{\alpha_{ij}} \ddot{h}_i \left(\log \frac{b_i}{y_i - r_i} \right) = \sum_{\substack{i \in \mathcal{I}_j \\ y_i \neq 0}} \frac{a_{ij}^2}{\alpha_{ij}} \frac{(y_i - r_i)^2}{y_i}. \quad (23)$$

The advantage of using this approximation is that one can precompute (23) *prior to iterating*. The accuracy of this approximation is illustrated in Fig. 2. To summarize, we replace the numerator of (20) with this approximation:

$$\begin{aligned} \frac{d}{d\theta_j} \phi_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^{\text{work}}} &\approx \frac{d}{d\theta_j} \hat{\phi}_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^{\text{work}}} = \\ \dot{L}_j(\theta^n) - \hat{d}_j \cdot (\theta_j^{\text{work}} - \theta_j^n) - \beta \sum_k w_{jk} \dot{\psi}(\theta_j^{\text{work}} - \theta_k^n). \end{aligned} \quad (24)$$

For the denominator of (20), note that

$$-\frac{d^2}{d\theta_j^2} \phi_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^n} = d_j(\theta^n) + \beta \sum_k w_{jk} \ddot{\psi}(\theta_j^n - \theta_k^n).$$

Since ψ has bounded curvature:

$$\ddot{\psi}(x) = \frac{1}{(1 + |x/\delta|)^2} \leq 1, \quad (25)$$

we replace the denominator of (20) with

$$-\frac{d^2}{d\theta_j^2} \phi_j(\theta_j; \theta^n) \Big|_{\theta_j = \theta_j^{\text{work}}} \approx \hat{d}_j + \beta \sum_k w_{jk}, \quad (26)$$

which is independent of θ^n so can be precomputed as described in [6]. Note that this replacement has no effect on the fixed point of (20). Using the approximation (26) provides a form of built-in under-relaxation because of the bounded curvature (25) of ψ .

To summarize, for our algorithm for performing the maximization (19), we replace (20) with (24) and (26), and apply 2 or 3 sub-iterations of the form (20). No forward or backprojections are computed during these sub-iterations, so they compute quickly. As in [5, 6, 10, 11], we store the current “forward projection” $\{\langle a_i, \theta^n \rangle\}$ to further save computation when evaluating the “backprojection” step (8). Since proper ordering of the steps is essential for efficient computation, we give the details of the algorithm in Table 1. (Software is also available; see [46].) The Appendix describes a modification to (29) that further improves the rate of convergence.

Precompute:

Initialize $\hat{\theta}$ via FBP

$$\hat{l}_i = \sum_{j=1}^p a_{ij} \hat{\theta}_j, \quad i = 1, \dots, N. \quad \hat{d}_j = \sum_{\substack{i \in \mathcal{I}_j \\ y_i \neq 0}} \frac{a_{ij}^2 (y_i - r_i)^2}{\alpha_{ij} y_i}, \quad \forall j.$$

for each iteration:

for each \mathcal{S} :

$$\dot{h}_i = \left[1 - \frac{y_i}{b_i e^{-\hat{l}_i} + r_i} \right] b_i e^{-\hat{l}_i}, \quad i = 1, \dots, N \quad (27)$$

for each $j \in \mathcal{S}$:

$$\dot{L}_j = \sum_{i \in \mathcal{I}_j} a_{ij} \dot{h}_i \quad (28)$$

$$\theta_j^{\text{work}} = \hat{\theta}_j$$

for a couple subiterations:

$$\theta_j^{\text{work}} := \left[\theta_j^{\text{work}} + \frac{\dot{L}_j - \hat{d}_j \cdot (\theta_j^{\text{work}} - \hat{\theta}_j) - \beta \sum_k w_{jk} \psi(\theta_j^{\text{work}} - \hat{\theta}_k)}{\hat{d}_j + \beta \sum_k w_{jk}} \right]_+ \quad (29)$$

end

$$\hat{l}_i := \hat{l}_i + a_{ij} (\theta_j^{\text{work}} - \hat{\theta}_j), \quad \forall i \text{ s.t. } a_{ij} \neq 0 \quad (30)$$

$$\hat{\theta}_j := \theta_j^{\text{work}}$$

end

end

end

Table 1: The general grouped-coordinate ascent algorithm. Note that the updates of $\hat{\theta}_j$ are done ‘‘in place.’’

IV. CONVERGENCE RATE AND ALGORITHM DESIGN

The method described in the preceding section is a class of algorithms since there are several factors that the algorithm designer may specify. Most importantly, one can choose the size and constituent elements of the groups \mathcal{S} for each n , as well as the factors α_{ij} (subject to (13)). The parameter β , the w_{jk} 's, and the function ψ are design choices too, but these determine the objective function, not the algorithm (at least within the class of convex ψ functions with bounded curvature). This section describes how the algorithm design factors influence convergence rate and computation time, starting with the α_{ij} 's.

If one were to use a single sub-iteration of the Newton-Raphson update⁹, then the ‘‘maximization step’’ ((19), (20), and (29)) would have the following form:

$$\theta_{\mathcal{S}}^{n+1} = \theta_{\mathcal{S}}^n + \mathbf{D}^{-1} \nabla_{\theta_{\mathcal{S}}}^T \Phi(\theta^n), \quad (31)$$

where \mathbf{D} is a $|\mathcal{S}| \times |\mathcal{S}|$ diagonal matrix with entries $\{\hat{d}_j + \beta \sum_k w_{jk}\}_{j \in \mathcal{S}}$. We could use (31) to develop expressions for the asymptotic convergence rate of the algorithm (for any particular choice of α_{ij} 's and \mathcal{S} 's) following the analysis in [23]. Here we take a more informal approach and simply note that (31) suggests that smaller values for the diagonal entries of \mathbf{D} will lead to larger step sizes, and hence faster convergence¹⁰.

A. Choosing α_{ij} 's

If the diagonal entries \hat{d}_j of \mathbf{D} are to be made small, then from (22) we want the α_{ij} 's to be as large as possible, but subject to the constraint (13). Clearly this constraint depends on one's choices for \mathcal{S} , but for the moment assume we have fixed \mathcal{S} and we want to choose the α_{ij} 's.

De Pierro [25] proposed an algorithm for emission tomography that updates all pixels simultaneously (i.e. $\mathcal{S} = \{1, \dots, p\}$) and essentially uses (12) with

$$\alpha_{ij} = \frac{a_{ij} \theta_j^n}{\sum_{k \in \mathcal{S}} a_{ik} \theta_k^n}. \quad (32)$$

This was also applied to transmission tomography in [20]. The choice (32) has three disadvantages. First, if $\theta_j^n = 0$, then $\alpha_{ij} = 0$, so (22) would not be well defined. This complicates both implementation and convergence analysis. Second, as $\theta_j^n \rightarrow 0$, $\alpha_{ij} \rightarrow 0$, so $\hat{d}_j \rightarrow \infty$. Thus, pixels that approach 0 in the limit will converge increasingly slowly, perhaps even at sub-linear rates (as observed in the emission case [45]). Third, the choice (32) makes \hat{d}_j dependent on θ^n , so \hat{d}_j cannot be precomputed.

One way to overcome the first two drawbacks is to express the emission algorithm (PML-SAGE-3) developed in [24] in terms of De Pierro's convexity method. This leads to the following choice:

$$\alpha_{ij} = \frac{a_{ij} (\theta_j^n + z_j)}{\sum_{k \in \mathcal{S}} a_{ik} (\theta_k^n + z_k)}, \quad (33)$$

⁹One sub-iteration is adequate when ψ is quadratic, for example, or when the algorithm has nearly converged. So (31) is useful for studying the asymptotic convergence rate.

¹⁰Excepting possible acceleration for small $|\mathcal{S}|$ due to under-relaxation as noted in [6, 24] for quadratic penalties.

for almost¹¹ any positive values z_j . Since z_j is positive, α_{ij} will be positive (when $a_{ij} \neq 0$). In results not shown, we have confirmed that this does lead to faster convergence than (32), presumably because the larger α_{ij} values lead to generally smaller \hat{d}_j values and hence larger step sizes.

However, the choice (33) still depends on θ^n , precluding precomputing \hat{d}_j . To eliminate this dependency, we let $z_j \rightarrow \infty$ in (33). This leads to the following choice:

$$\alpha_{ij} = \frac{a_{ij}}{\sum_{k \in \mathcal{S}} a_{ik}}, \quad (34)$$

which is independent of θ^n . This choice is similar to that made by De Pierro for the emission penalty function in [26], and was used in [27, 37]. Note that the denominator in (34) can be easily precomputed and stored once-and-for-all for a given tomographic system and choices for \mathcal{S} .

We use the choice (34) for the remainder of this paper. Whether better choices exist is an open question [47].

B. Special Cases

In the special case where the subset \mathcal{S} contains only one pixel ($\mathcal{S} = \{j\}$), then the ‘‘algorithm’’ (19) is equivalent to SCA [6, 10, 11], i.e., it turns out that

$$\phi_j(\theta_j; \theta^n) = \Phi(\theta_1^n, \dots, \theta_{j-1}^n, \theta_j, \theta_{j+1}^n, \dots, \theta_p^n).$$

And in that case, the choice (34) leads to a coordinate-wise Newton-Raphson update [6, 10, 11].

At the other extreme, when $\mathcal{S} = \{1, \dots, p\}$, then using the choice (32) with one sub-iteration of (20) is equivalent to the convex algorithm of [20]. The choice (34) thus corresponds to an alternative convex algorithm (and one that converges faster).

However, the algorithms that are ‘‘in between’’ those two extreme choices of \mathcal{S} are the most useful, as discussed next.

C. Choosing Groups \mathcal{S}

Optimization algorithms of the class described above seem to involve the following tradeoff. The more parameters one updates simultaneously, the smaller the step sizes must be to ensure monotonicity, since the parameters are coupled. Specifically, from (32), (33), and (34), as the size of \mathcal{S} increases, the α_{ij} values typically decrease, leading to larger \hat{d}_j 's and hence smaller step sizes in (31). So updating the parameters in smaller groups typically yields faster per-iteration convergence rates, with SCA (one parameter at a time) being the extreme case. However, as mentioned above there are often ‘‘economies of scale’’ that one can exploit when updating several parameters simultaneously. So the actual computation per iteration is often reduced by updating larger groups. Thus for fast convergence but moderate computation, we would like to update the parameters using a few large groups, but chosen such that the parameters within each group are relatively uncoupled. By uncoupled, we mean that the α_{ij} terms are not too much smaller than 1, which is the value that α_{ij} takes when $\mathcal{S} = \{j\}$. Specifically, note from (34) that the only i indices that matter are those in

¹¹The constraint (16) may need to be considered.

\mathcal{I}_j . If one can choose \mathcal{S} so that for $k \in \mathcal{S}$ the values of a_{ik} are small for $i \in \mathcal{I}_j$, then $\alpha_{ij} \approx 1$. For most tomographic geometries with finite-width rays and pixels, there is at least one ray that intersects any given pair of pixels, so one cannot simultaneously achieve $\alpha_{ij} = 1$ with multiple-pixel choices for \mathcal{S} . But pixels that are closer together typically share more intersecting rays than those that are well-separated spatially, so if \mathcal{S} contains only spatially well-separated pixels, the values for α_{ij} should be reasonably close to 1. (One might ask “why not just increase the step size in (31) using an over-relaxation parameter?” The danger is that such over-relaxation can destroy monotonicity.)

We have investigated the following GCA method. We divide the image into blocks of size $m \times m$, for small m , and then update only 1 pixel out of each $m \times m$ block on a given sub-iteration¹². The number of groups is thus m^2 , with p/m^2 pixels per group. Thus the required number of exponentiations is only $m^2 N$, which is considerably smaller than the number of nonzero a_{ij} ’s for small m . Note that $m = 1$ is closely related to the convex algorithm [20], and $m = \sqrt{p}$ gives the SCA algorithm [6]. As one increases m , the pixels within each group become more separated and therefore less coupled, which increases the convergence rate, but the computation also increases. Thus there is a basic tradeoff that can be adapted to the characteristics of the particular computer architecture.

Method	Water	Spine	Lung
FBP, 14 hour			
ROI Mean	0.0939	0.1662	0.0345
FBP, 12 min.			
ROI Mean	0.0942	0.1685	0.0373
ROI Std. Dev.	0.0098	0.0115	0.0068
PL-GCA, 12 min.			
ROI Mean	0.0945	0.1656	0.0353
ROI Std. Dev.	0.0030	0.0055	0.0015

Table 3: Mean and standard deviations within rectangular regions of interest for the images shown in Fig. 3.

V. RESULTS

In [27] we presented convergence rate results using simulated PET transmission scans. Here we present analogous results using real data. Using an Siemens/CTI ECAT EXACT 921 PET scanner equipped with rotating rod transmission sources [1], we acquired a 15-hour blank scan (b_i ’s) and two transmission scans (y_i ’s) of an anthropomorphic thorax phantom (Data Spectrum, North Carolina). The duration of one transmission scan was 14 hours (64M prompt coincidences in the slice of interest) and the other scan was 12 minutes (0.921M prompt coincidences in the slice of interest). (Most of these counts correspond to rays that do not intersect the object.) Delayed coincidence sinograms were collected separately. The blank and transmission scan delayed-coincidence sinograms were in close agreement, so we used a time-scaled version of the blank scan delayed coincidences as the r_i factors with no additional processing. The

¹²Similar “generalized checkerboard” decompositions of the image have been considered for emission tomography [48] [49].

sinogram dimension was 160 radial bins and 192 angles, and the reconstructed images were 128^2 with 4.5mm pixels. For the a_{ij} ’s, we used 6mm wide strip integrals having 3mm spacing [6], which roughly approximates the system geometry.

Reconstructions of the phantom are shown in Fig. 3, by both FBP and by 20 iterations of 4×4 GCA. For the penalized likelihood reconstructions we used $\delta = 0.004\text{cm}^{-1}$ in (6), chosen by visual inspection. The qualitative properties were rather sensitive to the choice of this parameter. (A 3D penalty function might reduce this sensitivity by improving the reconstruction of thin axial structures such as the patient table in Fig. 3.) The statistical method appears to produce somewhat better image quality. (See [6] for quantitative resolution-vs-noise comparisons.)

Fig. 4 shows that with $m = 4$ (16 groups), the proposed GCA algorithm increased the penalized-likelihood objective almost as fast as the SCA algorithm per iteration. More important is the actual CPU time, which is shown in Fig. 5 (for a DEC AlphaStation 600-5/266 workstation). By using fewer exponentiations and floating point operations, the GCA algorithms require far less CPU time per iteration than the SCA algorithm. Table 2 compares the number of iterations and CPU seconds required to (nearly) maximize the penalized-likelihood objective function Φ . With $m = 3$ or $m = 4$, the GCA algorithms converge in less than half the CPU time of SCA. Furthermore, the GCA algorithms are parallelizable, so with appropriate hardware could be significantly accelerated. Note that “ 1×1 GCA” is closely related to the convex algorithm of [20].

Table 3 compares the estimated attenuation coefficients for three rectangular regions of interest (ROIs) corresponding to soft tissue, bone, and lung. The ROI values for the 12-minute data both agree well with the 14-hour reference image. However, the within-ROI standard deviations for the penalized-likelihood image are factors of 2-4.5 smaller than those of the FBP image.

VI. DISCUSSION

We have described a new class of algorithms for maximizing (almost) concave penalized-likelihood objective functions for reconstructing attenuation images from low-count transmission scans. There is considerable latitude for the algorithm designer to choose algorithm parameters to achieve the fastest possible convergence rate on a given computer architecture. When the objective function is concave, the algorithm converges globally to the unique maximum. Thus the algorithm design parameters only affect the convergence rate, not the image quality, unlike the many popular unregularized methods.

Our results demonstrate that even on a conventional workstation the new algorithms converge faster than both SCA and (an improved version of) the convex algorithm of [20]. The results in [20] and [6] provide additional comparisons to other alternative algorithms. Based on all of these comparisons, we consider the transmission EM algorithm [9, 18] to be obsolete. For penalized-likelihood transmission image reconstruction, our proposed GCA algorithms have fast convergence, reduced exponentiations per iteration, easily accommodate non-negativity, and are flexibly parallelizable.

From Table 2, to process the 47 slices of an EXACT PET

	GCA				SCA
	1 × 1	2 × 2	3 × 3	4 × 4	
Number of iterations for convergence	>40	19	14	13	11
CPU seconds for convergence	>54	30	24	24	56
CPU seconds per iteration	1.2	1.3	1.4	1.5	4.8

Table 2: Comparison of CPU times and iterations for the proposed GCA algorithms versus the SCA algorithm. For purposes of this table, convergence means $\Phi(\theta^n) - \Phi(\theta^0) > 0.999[\Phi(\hat{\theta}) - \Phi(\theta^0)]$.

scanner using 14 iterations of 3×3 GCA requires about 19 minutes on a DEC AlphaStation (whereas SCA would require about 44 minutes). Such processing times bring this statistical method within the realm of clinical utility, although further time reductions would still be helpful.

One could combine the grouped-ascent idea in this paper with the hybrid Poisson/polynomial approximations described in [6] to further reduce computation. The reductions would be less dramatic than in [6] since for our GCA method the exponentiations in Table 1 have been moved outside of the backprojection step, whereas for SCA the calculations in (27) must be done during the backprojection (8) since θ^n is continually changing.

There are additional advantages of GCA that we have not exploited here. Relative to SCA, which is best suited to a_{ij} 's that are precomputed and stored, the GCA approach can more easily exploit the many tricks available for accelerating forward- and back-projection operations ((28), (30)), such as symmetries in the a_{ij} 's, projection operators based on image rotation, and a_{ij} 's that separate into sparse line-integrals following by a space-invariant blur implemented using fast Fourier transforms. In some applications these tricks should lead to further reductions in computation time. Additional improvements may follow from further algorithm development. A natural starting point would be to relax the separability assumption (11).

VII. APPENDIX

This appendix presents a method for finding the zero-crossing of $d/d\theta_j \hat{\phi}_j(\cdot; \theta^n)$ as defined by (24). This method converges faster than the modified Newton-Raphson method given in Table 1. Define $x = \theta_j - \theta_j^n$, and $x_k = \theta_k^n - \theta_j^n$, and

$$g(x) = \dot{L}_j(\theta^n) - \hat{d}_j \cdot x - \beta \sum_k w_{jk} \dot{\psi}(x - x_k),$$

so that $g(x) = d/d\theta_j \hat{\phi}_j(\theta_j^n + x; \theta^n)$. We would like to find the value \hat{x} where $g(\hat{x}) = 0$ (i.e. its zero crossing), and then assign $\theta_j^{n+1} = \theta_j^n + \hat{x}$. Let N_j be the number of neighbors of pixel j , i.e., the number of nonzero w_{jk} terms (typically $N_j = 8$). Observe that $g(x)$ is the sum of $1 + N_j$ monotonically decreasing functions; the first of these functions is $\dot{L}_j(\theta^n) - \hat{d}_j \cdot x$, which crosses zero at $x = \dot{L}_j(\theta^n)/\hat{d}_j$, and the other N_j functions are the penalty terms, the k th of which crosses zero at $x = x_k$.

The zero-crossing of $g(x)$ must occur somewhere between the maximum and minimum of those $1 + N_j$ individual zero

crossings¹³. We first search over that set of $1 + N_j$ candidate zero crossings (we also check the values 0 and $\pm\theta_j^n \varepsilon$ where ε is about 0.02) to bound the zero crossing of $g(x)$ within an interval (x_-, x_+) . Although the curvature of ψ is certainly bounded above by 1 as described in (25) and (26), its curvature is bounded above by an even smaller value over the interval (x_-, x_+) . Specifically:

$$\ddot{\psi}(x - x_k) \leq \max\{\ddot{\psi}(x_- - x_k), \ddot{\psi}(x_+ - x_k)\} \triangleq \gamma_k. \quad (35)$$

Note that $\gamma_k \leq 1$. Thus we replace the denominator in (29) with

$$\hat{d}_j + \beta \sum_k w_{jk} \gamma_k.$$

This leads to faster convergence since the denominator in (29) is smaller, therefore the step size is larger. Note that by using the bound in (35) rather than some ad hoc value, we still ensure monotonic increases in $\hat{\phi}_j(\cdot; \theta^n)$.

REFERENCES

- [1] K Wienhard, L Eriksson, S Grootoink, M Casey, U Pietrzyk, and W D Heiss. Performance evaluation of a new generation positron scanner ECAT EXACT. *J. Comp. Assisted Tomo.*, 16(5):804–813, Sept. 1992.
- [2] S R Cherry, M Dahlbom, and E J Hoffman. High sensitivity, total body PET scanning using 3D data acquisition and reconstruction. *IEEE Tr. Nuc. Sci.*, 39(4):1088–1092, Aug. 1992.
- [3] E P Ficaro, J A Fessler, W L Rogers, and M Schwaiger. Comparison of Americium-241 and Technicium-99m as transmission sources for attenuation correction of Thallium-201 SPECT imaging of the heart. *J. Nuc. Med.*, 35(4):652–663, Apr. 1994.
- [4] S R Meikle, M Dahlbom, and S R Cherry. Attenuation correction using count-limited transmission data in positron emission tomography. *J. Nuc. Med.*, 34(1):143–150, Jan. 1993.
- [5] K Sauer and C Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Tr. Sig. Proc.*, 41(2):534–548, Feb. 1993.

¹³Thanks to Ken Sauer for bringing this point to the attention of the first author when discussing [11].

- [6] J A Fessler. Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans. *IEEE Tr. Im. Proc.*, 4(10):1439–50, Oct. 1995.
- [7] D S Lalush and B M W Tsui. MAP-EM and WLS-MAP-CG reconstruction methods for transmission imaging in cardiac SPECT. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1174–1178, 1993.
- [8] J A Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Tr. Im. Proc.*, 5(3):493–506, Mar. 1996.
- [9] K Lange and R Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–316, Apr. 1984.
- [10] C Bouman and K Sauer. Fast numerical methods for emission and transmission tomographic reconstruction. In *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, pp. 611–616, 1993.
- [11] C A Bouman and K Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Tr. Im. Proc.*, 5(3):480–92, Mar. 1996.
- [12] J A Fessler and W L Rogers. Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs. *IEEE Tr. Im. Proc.*, 5(9):1346–58, Sept. 1996.
- [13] A J Rockmore and A Macovski. A maximum likelihood approach to transmission image reconstruction from projections. *IEEE Tr. Nuc. Sci.*, 24(3):1929–1935, June 1977.
- [14] K Lange, M Bahn, and R Little. A theoretical study of some maximum likelihood algorithms for emission and transmission tomography. *IEEE Tr. Med. Im.*, 6(2):106–114, June 1987.
- [15] K Lange. An overview of Bayesian methods in image reconstruction. In *Proc. SPIE 1351, Dig. Im. Synth. and Inverse Optics*, pp. 270–287, 1990.
- [16] K Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Tr. Med. Im.*, 9(4):439–446, Dec. 1990. Corrections, June 1991.
- [17] E Mumcuoglu, R Leahy, and S Cherry. A statistical approach to transmission image reconstruction from ring source calibration measurements in PET. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 910–912, 1992.
- [18] J M Ollinger. Maximum likelihood reconstruction of transmission images in emission computed tomography via the EM algorithm. *IEEE Tr. Med. Im.*, 13(1):89–101, Mar. 1994.
- [19] E U Mumcuoglu, R Leahy, S R Cherry, and Z Zhou. Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images. *IEEE Tr. Med. Im.*, 13(3):687–701, Dec. 1994.
- [20] K Lange and J A Fessler. Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Tr. Im. Proc.*, 4(10):1430–8, Oct. 1995.
- [21] E Ü Mumcuoğlu and R M Leahy. A gradient projection conjugate gradient algorithm for Bayesian PET reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 1212–6, 1994.
- [22] L Kaufman. Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Tr. Med. Im.*, 12(2):200–214, June 1993.
- [23] J A Fessler and A O Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Tr. Sig. Proc.*, 42(10):2664–2677, Oct. 1994.
- [24] J A Fessler and A O Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Tr. Im. Proc.*, 4(10):1417–29, Oct. 1995.
- [25] A R De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Tr. Med. Im.*, 12(2):328–333, June 1993.
- [26] A R De Pierro. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Tr. Med. Im.*, 14(1):132–137, Mar. 1995.
- [27] J A Fessler, E P Ficaro, N H Clinthorne, and K Lange. Fast parallelizable algorithms for transmission image reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 1346–50, 1995.
- [28] G Gullberg and B M W Tsui. Maximum entropy reconstruction with constraints: iterative algorithms for solving the primal and dual programs. In C N de Graaf and M A Viergever, editors, *Proc. Tenth Intl. Conf. on Information Processing in Medical Im.*, pp. 181–200. Plenum Press, New York, 1987.
- [29] J A Fessler. Penalized weighted least-squares image reconstruction for positron emission tomography. *IEEE Tr. Med. Im.*, 13(2):290–300, June 1994.
- [30] M Wax. *Detection and estimation of superimposed signals*. PhD thesis, Stanford Univ., Stanford, CA., Mar. 1985.
- [31] I Ziskind and M Wax. Maximum likelihood localization of multiple sources by alternating projection. *IEEE Tr. Acoust. Sp. Sig. Proc.*, 36(10):1553–1560, Oct. 1988.
- [32] M Feder and E Weinstein. Parameter estimation of superimposed signals using the EM algorithm. *IEEE Tr. Acoust. Sp. Sig. Proc.*, 36(4):477–489, Apr. 1988.
- [33] A J Weiss, A S Willsky, and B C Levy. Maximum likelihood array processing for the estimation of superimposed signals. *Proc. IEEE*, 76(2):203–205, Feb. 1988. Correction in *IEEE Proc.*, Vol. 76, No. 6, p. 734, June 1988.

- [34] J A Fessler. *Object-Based 3-D Reconstruction of Arterial Trees from a Few Projections*. PhD thesis, Stanford Univ., Stanford, CA., Aug. 1990.
- [35] J Besag. On the statistical analysis of dirty pictures. *J. Royal Stat. Soc. Ser. B*, 48(3):259–302, 1986.
- [36] J A Cadzow. Signal processing via least squares error modeling. *IEEE Signal Proc. Mag.*, pp. 12–31, Oct. 1990.
- [37] K D Sauer, S Borman, and C A Bouman. Parallel computation of sequential pixel updates in statistical tomographic reconstruction. In *Proc. IEEE Intl. Conf. on Image Processing*, vol. 3, pp. 93–6, 1995.
- [38] S T Jensen, S Johansen, and S L Lauritzen. Globally convergent algorithms for maximizing a likelihood function. *Biometrika*, 78(4):867–77, 1991.
- [39] M E Casey and E J Hoffman. Quantitation in positron emission computed tomography: 7 A technique to reduce noise in accidental coincidence measurements and coincidence efficiency calibration. *J. Comp. Assisted Tomo.*, 10(5):845–850, 1986.
- [40] B Chan, M Bergström, M R Palmer, C Sayre, and B D Pate. Scatter distribution in transmission measurements with positron emission tomography. *J. Comp. Assisted Tomo.*, 10(2):296–301, Mar. 1986.
- [41] G T Herman. *Image reconstruction from projections: The fundamentals of computerized tomography*. Academic Press, New York, 1980.
- [42] J M M Anderson, B A Mair, M Rao, and C H Wu. A weighted least-squares method for PET. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1292–6, 1995.
- [43] J A Fessler. Resolution properties of regularized image reconstruction methods. Technical Report 297, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Aug. 1995.
- [44] J A Fessler and W L Rogers. Uniform quadratic penalties cause nonuniform image resolution (and sometimes vice versa). In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 4, pp. 1915–1919, 1994.
- [45] J A Fessler, N H Clinthorne, and W L Rogers. On complete data spaces for PET reconstruction algorithms. *IEEE Tr. Nuc. Sci.*, 40(4):1055–1061, Aug. 1993.
- [46] J A Fessler. ASPIRE 3.0 user’s guide: A sparse iterative reconstruction library. Technical Report 293, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, July 1995. Available from WWW at <http://www.eecs.umich.edu/~fessler/>.
- [47] K Lange. Numerical analysis for statisticians, 1996. Preprint of text.
- [48] A R De Pierro. A generalization of the EM algorithm for maximum likelihood estimates from incomplete data. Technical Report MIPG119, Med. Im. Proc. Group, Dept. of Radiol., Univ. of Pennsylvania, Feb. 1987.
- [49] J A O’Sullivan. Roughness penalties on finite domains. *IEEE Tr. Im. Proc.*, 4(9):1258–68, Sept. 1995.

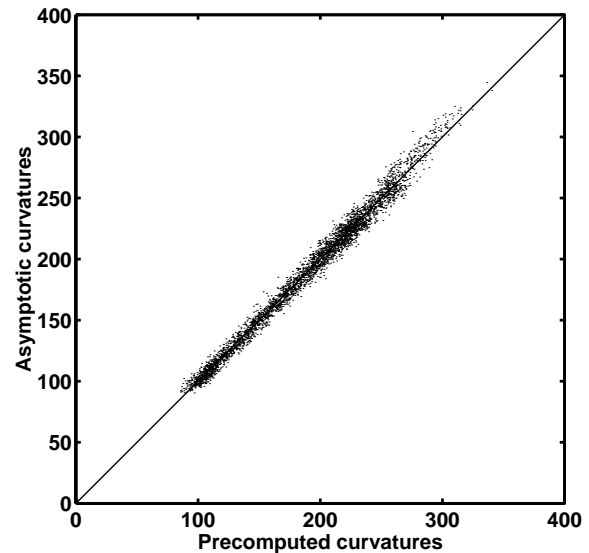


Figure 2: Comparison of the precomputed curvatures \hat{d}_j from (23), with the asymptotic curvatures $d_j(\hat{\theta})$, where $\hat{\theta}$ was taken to be the image shown at the bottom of Fig. 3. The precomputed values are very good approximations to the final curvatures of the Q_j surrogate functions.

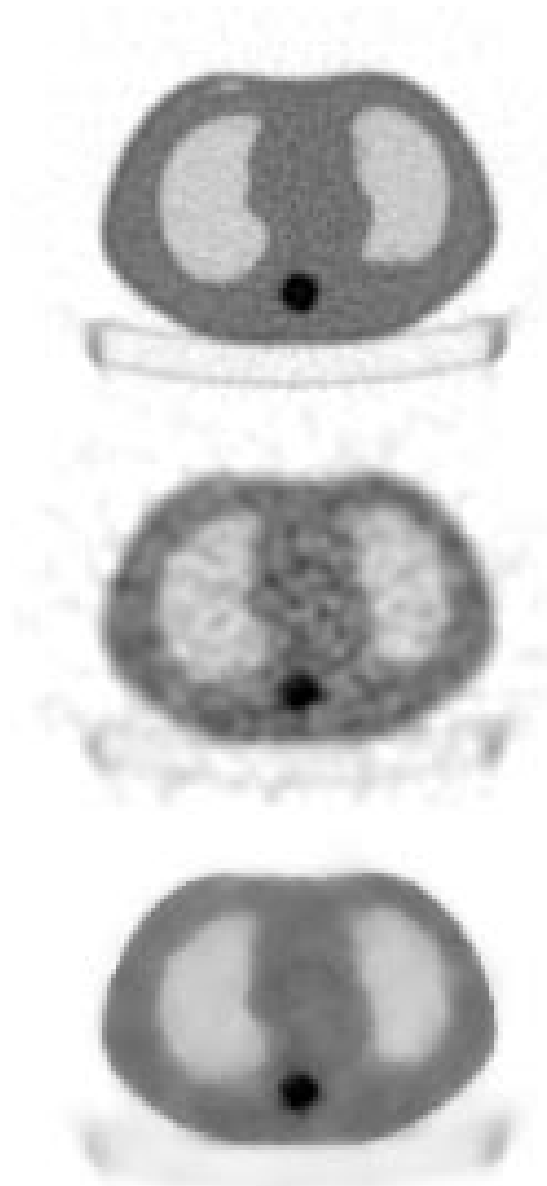


Figure 3: Top: FBP reconstruction of phantom data from 14 hour transmission scan. Middle: FBP reconstruction from 12 minute transmission scan. Bottom: Penalized-likelihood reconstruction from 12 minute transmission scan using 20 iterations of the 4×4 grouped coordinate ascent algorithm.

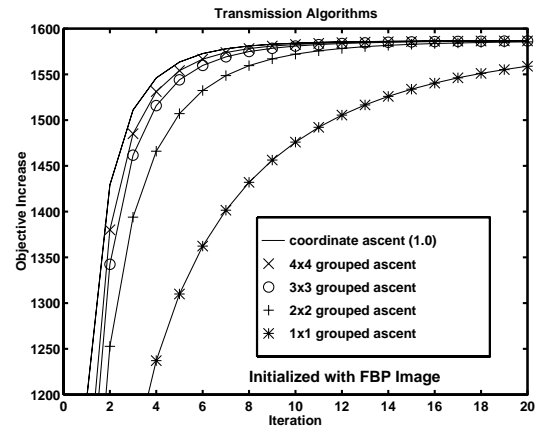


Figure 4: Objective function increase $\Phi(\theta^n) - \Phi(\theta^0)$ versus iteration n .

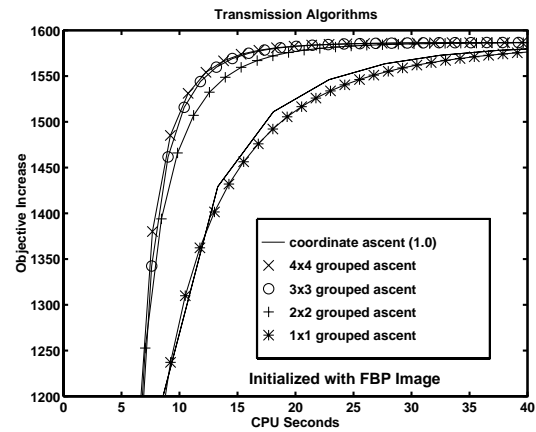


Figure 5: Objective function increase $\Phi(\theta^n) - \Phi(\theta^0)$ versus CPU time on a DEC AlphaStation.