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

Jeffrey A. Fessler,* *Member, IEEE*, Edward P. Ficaro, *Member, IEEE*, Neal H. Clinthorne, *Member, IEEE*, and Kenneth Lange

***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 maximum likelihood-expectation maximization (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.**

***Index Terms*— Biomedical nuclear imaging, Gauss–Seidel method, iterative methods, maximum likelihood estimation, nuclear tomography, positron emission tomography, single photon emission computed tomography.**

## I. Introduction

STATISTICAL methods for reconstructing attenuation images from transmission scans have increased in importance recently for several reasons, including the necessity of reconstructing two-dimensional (2-D) attenuation maps for reprojection to form three-dimensional (3-D) attenuation correction factors in septaless positron emission tomography (PET) [1], [2], the widening availability of single photon emission computed tomography (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 nonstatistical filtered backprojection (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 [9]–[11], [6]. 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

Manuscript received April 12, 1996; revised October 9, 1996. This work was supported in part by the National Institutes of Health under Grants CA-60711 and CA-54362. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was R. Leahy. *Asterisk indicates corresponding author*.

*J. A. Fessler is with the University of Michigan, 4240 EECS Bldg., Ann Arbor, MI 48109-2122 USA (email: fessler@umich.edu).

E. P. Ficaro, N. H. Clinthorne, and K. Lange are with the University of Michigan, Ann Arbor, MI 48109-2122 USA.

Publisher Item Identifier S 0278-0062(97)02405-1.

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 asymptotic convergence rate of the sequential AM algorithm relative to the simultaneous EM algorithm was shown in [34, ch. 5] (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 parallelizibility 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] and [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 (GCA) 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

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

where $N$ is the number of measurements, the inner product

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

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, $\theta_j$ 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 nonnegative constants.

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

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

where (neglecting constants independent of $\theta$ hereafter)

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

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 $\theta^{\text{true}}$, with $\hat{\theta}$ defined by

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

where the objective includes a roughness penalty

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

The function $\psi$ 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], and [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]

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

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

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

implies $|\dot{\psi}(x)| < \delta$, this potential function has bounded influence. The derivative (7) of $\psi$ 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 $\mathbb{R}^{\mathrm{P}}$. Since $\psi$ is strictly convex and $L(\cdot)$ is concave, the objective $\Phi$ 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 $\Phi$, nor do they easily accommodate nonnegativity constraints. Thus, for fastest convergence one must seek algorithms tailored to this problem. Relevant properties of $L$ are as follows.

- $L(\theta)$ 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

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

where

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

(see [6, (8)]) and where

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

An SCA algorithm must repeatedly evaluate $\dot{L}_j(\theta^n)$ at the current image estimate $\theta^n$. 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 simultaneously.[1] 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 algorithm p. 171). 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.

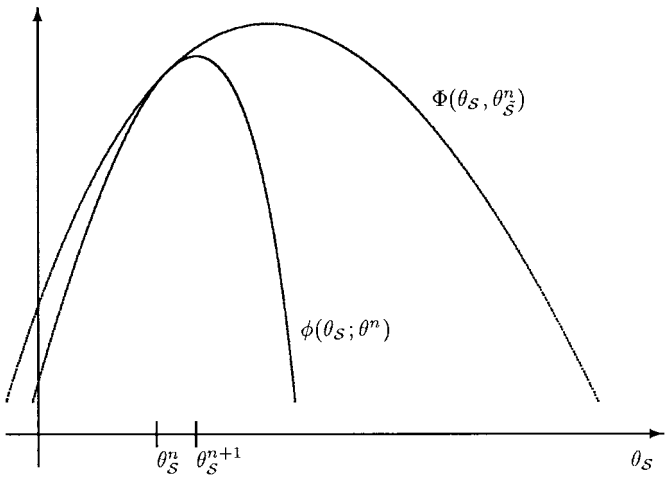


Fig. 1. One-dimensional (1-D) illustration of the optimization transfer principle: Instead of maximizing $\Phi(\theta_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}})$ over $\theta_{\mathcal{S}}$,we maximize the surrogate function $\phi(\theta_{\mathcal{S}}, \theta^n)$ iteratively. The higher the curvature of $\phi(\cdot, \theta^n)$, i.e., the greater the norm of its Hessian, the slower the convergence rate [23], [45].

Let $\mathcal{S}$ be a subset of the pixels $\{1, \cdots, p\}$, $\tilde{\mathcal{S}}$ be its complement, and $|\mathcal{S}|$ be the cardinality of $\mathcal{S}$. In a GCA algorithm,[2] we update $\theta_{\mathcal{S}}$ while holding $\theta^n_{\tilde{\mathcal{S}}}$ fixed at the $n$th update [23]. Unfortunately, it is even too difficult to maximize $\Phi(\theta_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}})$ over $\theta_{\mathcal{S}}$ directly, so we will settle for finding a method for choosing $\theta^{n+1}_{\mathcal{S}}$ that will at least *monotonically increase* the objective function[3]

$$\Phi\left(\theta^{n+1}_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}}\right) \geq \Phi\left(\theta^n_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}}\right) = \Phi(\theta^n).$$

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 $\theta^{n+1}_{\mathcal{S}}$ to maximize $\Phi(\theta_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}})$, 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

$$\Phi\left(\theta_{\mathcal{S}}, \theta^n_{\tilde{\mathcal{S}}}\right) - \Phi(\theta^n) \geq \phi\left(\theta_{\mathcal{S}}; \theta^n\right) - \phi\left(\theta^n_{\mathcal{S}}; \theta^n\right), \ \forall \theta_{\mathcal{S}} \in \mathcal{R}_{\mathcal{S}}(\theta^n). \qquad (9)$$

The GCA update (cf. space-alternating generalized EM (SAGE) algorithm [23], [24]) is then

$$\theta^{n+1}_{\mathcal{S}} = \arg \max_{\theta_{\mathcal{S}} \in \mathcal{R}_{\mathcal{S}}(\theta^n)} \phi(\theta_{\mathcal{S}}; \theta^n)$$
$$\theta^{n+1}_j = \theta^n_j, \quad j \in \tilde{\mathcal{S}}. \qquad (10)$$

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

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

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

[3] To 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.

### A. Choosing Surrogate Functions

We restrict attention here to additively separable[4] surrogate functions $\phi(\cdot;\theta^n)$ satisfying

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

To choose these $\phi_j$'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

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

for any choice[5] of $\alpha_{ij} \geq 0$ that satisfies the constraint

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

We discuss specific choices for $\alpha_{ij}$ in the next section.

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

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

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 $\theta_{\mathcal{S}}$ such that

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

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

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

where

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

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

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

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

for $\theta_{\mathcal{S}} \in \mathcal{R}_{\mathcal{S}}(\theta^n)$, where using (12) and (14)

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

[4] Separable surrogate functions are very convenient for enforcing the nonnegativity constraint. There may be alternatives that lead to faster convergence though.

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

Assuming the groups are chosen so that no two neighboring pixels are in the same group,[6] then the surrogate function defined by (11) with[7]

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

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

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

### B. Convergence

When $r_i = 0\ \forall i$ so that $\Phi$ 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 $\Phi$ and converges from any starting image to the unique global maximizer of $\Phi$ subject to $\theta \geq 0$, under mild assumptions about $\mathcal{S}$ and the $\alpha_{ij}$'s. There are a few practical caveats that should 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 1-D 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 $\theta^n_{j,\max}$ terms, so in our software we take the more pragmatic approach of simply verifying that $\Phi$ 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 nonconcave 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 subiterations of the 1-D Newton–Raphson method

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

where $[x]_+ = x$ for $x > 0$ and is zero, otherwise. This $[\cdot]_+$ operator enforces the nonnegativity constraint. The ":=" 

[6] If 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 two following the parameter $\beta$ in (26) and in the denominator of (29) [6].

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

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 $\theta_j^n$, 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 $\theta_j^n$ 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

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

and that

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

where (see [6])

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

Note that $\theta^n$ only enters $d_j(\theta^n)$ through its projections $\langle a_i, \theta^n\rangle$. Thus, $d_j(\theta^n)$ is fairly insensitive to $\theta^n$, 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

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

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

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

For the denominator of (20), note that

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

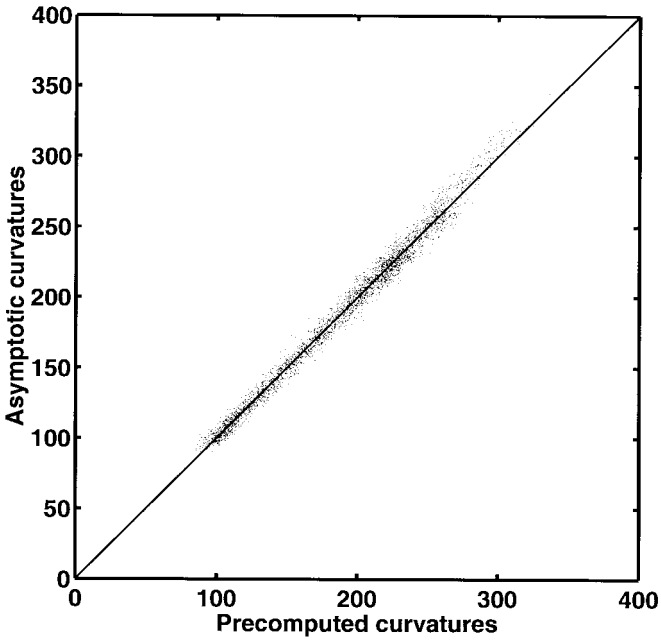


Fig. 2. Comparison of the precomputed curvatures $\hat{d}_j$ from (23) with the asymptotic curvatures $d_j(\hat{\theta})$, where $\hat{\theta}$ was taken to be the image shown at the bottom of Fig. 3. The precomputed values are very good approximations to the final curvatures of the $Q_j$ surrogate functions.

Since $\psi$ has bounded curvature

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

we replace the denominator of (20) with

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

which is independent of $\theta^n$, 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 $\psi$.

To summarize, for our algorithm for performing the maximization (19), we replace (20) with (24) and (26), and apply two or three subiterations of the form (20). No forward or backprojections are computed during these subiterations, so they compute quickly. As in [5], [6], [10], and [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 shown at the top of the next page. (Software is also available; see [46].) The Appendix describes a modification to (29) that further improves the rate of convergence.

## 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 $n$,

Precompute: Initialize $\hat{\theta}$ via FBP

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

`for` each iteration:

`for` each $S$:

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

`for` each $j \in S$:

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

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

`for` a couple subiterations:

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

`end`

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

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

`end`

`end`

`end`

as well as the factors $\alpha_{ij}$ [subject to (13)]. The parameter $\beta$, the $w_{jk}$'s, and the function $\psi$ are design choices too, but these determine the objective function, not the algorithm (at least within the class of convex $\psi$ functions with bounded curvature). This section describes how the algorithm design factors influence convergence rate and computation time, starting with the $\alpha_{ij}$'s.

If one were to use a single subiteration of the Newton–Raphson update,[8] then the "maximization step" [(19), (20), (29)] would have the following form:

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

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 $\alpha_{ij}$'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 convergence.[9]

[8] One subiteration is adequate when $\psi$ is quadratic, for example, or when the algorithm has nearly converged. So, (31) is useful for studying the asymptotic convergence rate.

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

### *A. Choosing $\alpha_{ij}$'s*

If the diagonal entries $\hat{d}_j$ of $\mathbf{D}$ are to be made small, then from (22) we want the $\alpha_{ij}$'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 $\alpha_{ij}$'s.

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

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

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 $\theta_j^n \to 0$, $\alpha_{ij} \to 0$, so $\hat{d}_j \to \infty$. Thus, pixels that approach zero in the limit will converge increasingly slowly, perhaps even at sublinear rates (as observed in the emission case [45]). Third, the choice (32) makes $\hat{d}_j$ dependent on $\theta^n$, 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:

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

for almost[10] any positive values $z_j$. Since $z_j$ is positive, $\alpha_{ij}$ 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 $\alpha_{ij}$ values lead to generally smaller $\hat{d}_j$ values and hence larger step sizes.

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

$$\alpha_{ij} = \frac{a_{ij}}{\sum_{k \in \mathcal{S}} a_{ik}} \tag{34}$$

which is independent of $\theta^n$. 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

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

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

At the other extreme, when $\mathcal{S} = \{1, \cdots, p\}$, then using the choice (32) with one subiteration 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)–(34), as the size of $\mathcal{S}$ increases, the $\alpha_{ij}$ 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 $\alpha_{ij}$ terms are not too much smaller than 1.0, which is the value that $\alpha_{ij}$ takes when $\mathcal{S} = \{j\}$. Specifically, note from (34) that the only $i$ indexes that matter are those in $\mathcal{I}_j$. 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 $\alpha_{ij}$ should be reasonably close to 1.0. (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 one pixel out of each $m \times m$ block on a given subiteration.[11] 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.

## 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-h blank scan ($b_i$'s) and two transmission scans ($y_i$'s) of an anthropomorphic thorax phantom (Data Spectrum, Chapel Hill, NC). The duration of one transmission scan was 14 h (64M prompt coincidences in the slice of interest) and the other scan was 12 min (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 sinogram dimension was 160 radial bins and 192 angles, and the reconstructed images were $128^2$ with 4.5-mm pixels. For the $a_{ij}$'s, we used 6-mm-wide strip integrals having 3-mm spacing [6], which roughly approximates the system geometry.

[10] The constraint (16) may need to be considered.

[11] Similar "generalized checkerboard" decompositions of the image have been considered for emission tomography [48], [49].

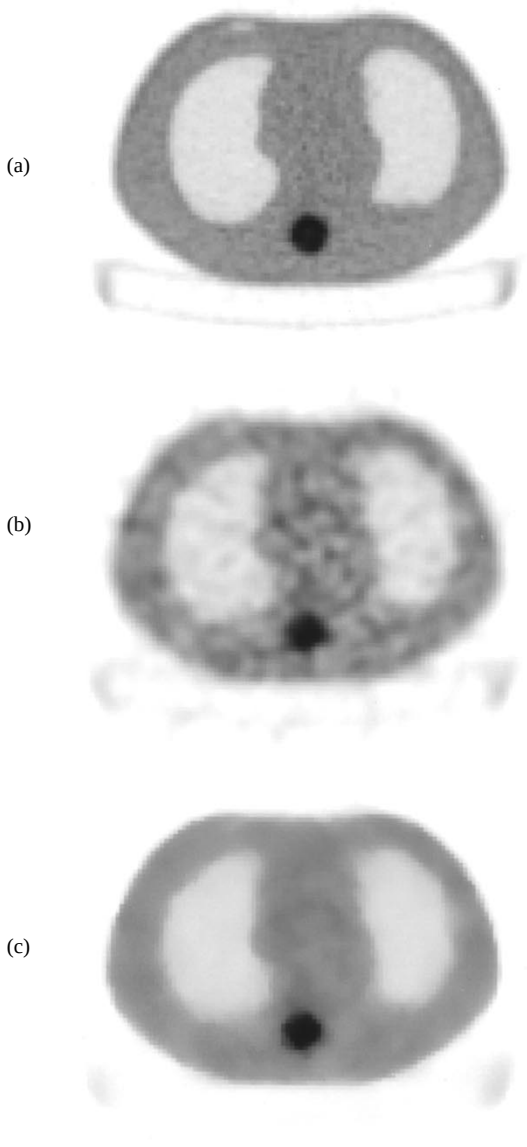


Fig. 3. (a) FBP reconstruction of phantom data from 14-h transmission scan, (b) FBP reconstruction from 12-min transmission scan, and (c) penalized-likelihood reconstruction from 12-min transmission scan using 20 iterations of the $4 \times 4$ GCA algorithm.

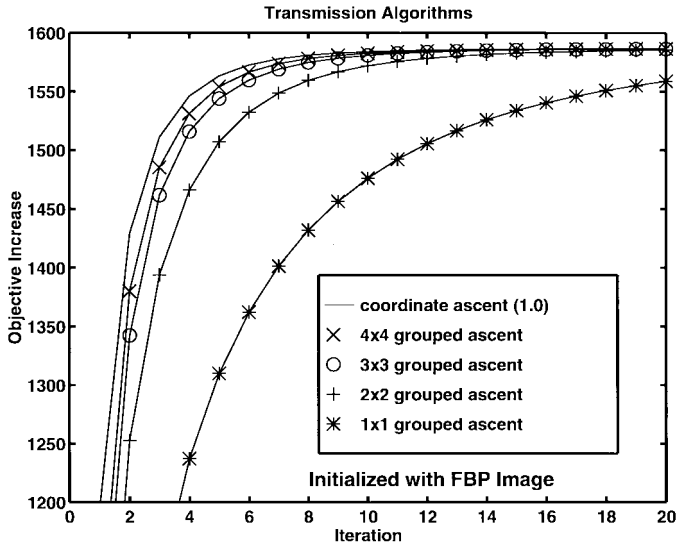


Fig. 4. Objective function increase $\Phi(\theta^n) - \Phi(\theta^0)$ versus iteration $n$.

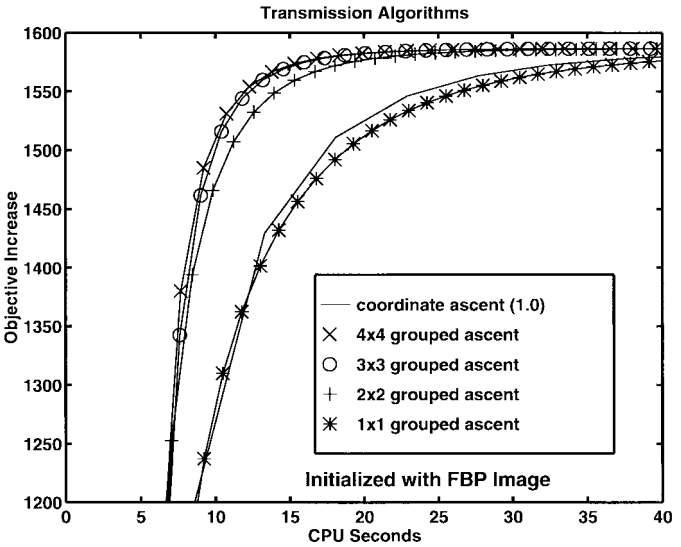


Fig. 5. Objective function increase $\Phi(\theta^n) - \Phi(\theta^0)$ versus CPU time on a DEC AlphaStation.

Reconstructions of the phantom are shown in Fig. 3, by both FBP and by 20 iterations of $4 \times 4$ GCA. For the penalized likelihood reconstructions we used $\delta = 0.004$ cm$^{-1}$ in (6), chosen by visual inspection. The qualitative properties were rather sensitive to the choice of this parameter. (A 3-D 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 versus 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 I compares the number of iterations and CPU seconds required to (nearly) maximize the penalized-likelihood objective function $\Phi$. 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 "$1 \times 1$ GCA" is closely related to the convex algorithm of [20].

Table II compares the estimated attenuation coefficients for three rectangular regions of interest (ROI's) corresponding to soft tissue, bone, and lung. The ROI values for the 12-min data both agree well with the 14-h 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.

TABLE I
COMPARISON OF CPU TIMES AND ITERATIONS FOR THE PROPOSED GCA ALGORITHMS VERSUS THE SCA ALGORITHM. FOR PURPOSES OF THIS TABLE, CONVERGENCE MEANS $\Phi(\theta^n) - \Phi(\theta^0) > 0.999[\Phi(\hat{\theta}) - \Phi(\theta^0)]$

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

TABLE II
MEAN AND STANDARD DEVIATIONS WITHIN RECTANGULAR REGIONS OF INTEREST FOR THE IMAGES SHOWN IN Fig. 3

| Method | | Water | Spine | Lung |
|---|---|---|---|---|
| FBP, 14-h | 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 |

## 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 [6] and [20] 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 nonnegativity, and are flexibly parallelizable.

From Table I, to process the 47 slices of an EXACT PET scanner using 14 iterations of $3 \times 3$ GCA requires about 19 min on a DEC AlphaStation (whereas SCA would require about 44 min). 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 the algorithm (See p. 171) have been moved outside of the backprojection step, whereas for SCA the calculations in (27) must be done during the backprojection (8) since $\theta^n$ 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 backprojection 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).

## 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 the subiteration (29). 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 crossings.[12] We first search over that set of $1 + N_j$ candidate zero crossings (we also check the values 0.0 and $\pm\theta_j^n \varepsilon$ where $\varepsilon$ is about 0.02) to bound the zero crossing of $g(x)$ within an interval $(x_-, x_+)$. Although the curvature of $\psi$ is certainly bounded above by 1.0 as described in (25) and (26), its curvature is bounded above by an even smaller value over the interval $(x_-, x_+)$. Specifically

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

Note that $\gamma_k \le 1$. 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 $\hat{\phi}_j(\cdot;\theta^n)$.

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

## REFERENCES


[1] K. Wienhard, L. Eriksson, S. Grootoonk, M. Casey, U. Pietrzyk, and W. D Heiss, "Performance evaluation of a new generation positron scanner ECAT EXACT," *J. Comput. Assist. Tomogr.*, vol. 16, no. 5, pp. 804–813, Sept. 1992.

[2] S. R Cherry, M. Dahlbom, and E. J Hoffman, "High sensitivity, total body PET scanning using 3-D data acquisition and reconstruction," *IEEE Trans. Nucl. Sci.*, vol. 39, no. 4, pp. 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. Nucl. Med.*, vol. 35, no. 4, pp. 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. Nucl. Med.*, vol. 34, no. 1, pp. 143–150, Jan. 1993.

[5] K. Sauer and C. Bouman, "A local update strategy for iterative reconstruction from projections," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 534–548, Feb. 1993.

[6] J. A. Fessler, "Hybrid poisson/polynomial objective functions for tomographic image reconstruction from transmission scans," *IEEE Trans. Image. Processing*, Oct. 1995, vol. 4, no. 10, pp. 1439–50.

[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 Nucl. Sci. Symp. Med. Imag. Conf.*, 1993, vol. 2, pp. 1174–1178.

[8] J. A. Fessler, "Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. Image. Processing*, Mar. 1996, vol. 5, no. 3, pp. 493–506.

[9] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comput. Assist. Tomogr.*, vol. 8, no. 2, pp. 306–316, Apr. 1984.

[10] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. 27th Conf. Inform. Sci. Syst.*, Johns Hopkins, 1993, pp. 611–616.

[11] ———, "A unified approach to statistical tomography using coordinate descent optimization," *IEEE Trans. Image Processing*, Mar. 1996, vol. 5, no. 3, pp. 480–92.

[12] J. A. Fessler and W. L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs," *IEEE Trans. Image Processing*, Sept. 1996, vol. 5, no. 9, pp. 1346–58.

[13] A. J. Rockmore and A. Macovski, "A maximum likelihood approach to transmission image reconstruction from projections," *IEEE Trans. Nucl. Sci.*, vol. 24, no. 3, pp. 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 Trans. Med. Imag.*, vol. MI-6, no. 2, pp. 106–114, June 1987.

[15] K. Lange, "An overview of Bayesian methods in image reconstruction," in *Proc. SPIE 1351, Dig. Image Synth. and Inverse Optics*, 1990, pp. 270–287.

[16] ———, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, no. 4, pp. 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 Nucl. Sci. Symp. Med. Imag. Conf.*, 1992, vol. 2, pp. 910–912.

[18] J. M. Ollinger, "Maximum likelihood reconstruction of transmission images in emission computed tomography via the EM algorithm," *IEEE Trans. Med. Imag.*, vol. 13, no. 1, pp. 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 Trans. Med. Imag.*, vol. 13, no. 3, pp. 687–701, Dec. 1994.

[20] K. Lange and J. A. Fessler, "Globally convergent algorithms for maximum a posteriori transmission tomography," *IEEE Trans. Image Processing*, Oct. 1995, vol. 4, no. 10, pp. 1430–8.

[21] E. Ü. Mumcuoğlu and R. M. Leahy, "A gradient projection conjugate gradient algorithm for Bayesian PET reconstruction," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf.*, 1994, vol. 3, pp. 1212–6.

[22] L. Kaufman, "Maximum likelihood, least squares, and penalized least squares for PET," *IEEE Trans. Med. Imag.*, vol. 12, no. 2, pp. 200–214, June 1993.

[23] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, no. 10, pp. 2664–2677, Oct. 1994.

[24] ———, "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Trans. Image Processing*, Oct. 1995, vol. 4, no. 10, pp. 1417–29.

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

[26] ———, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Trans. Med. Imag.*, vol. 14, no. 1, pp. 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 Nucl. Sci. Symp. Med. Imag. Conf.*, 1995, vol. 3, pp. 1346–1350.

[28] G. Gullberg and B. M. W. Tsui, "Maximum entropy reconstruction with constraints: Iterative algorithms for solving the primal and dual programs," in *Proc. 10th Int. Conf. Information Processing in Medical Imag.*, C. N. de Graaf and M. A. Viergever, Eds.) New York: Plenum, 1987, pp. 181–200.

[29] J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. 13, no. 2, pp. 290–300, June 1994.

[30] M. Wax, "Detection and estimation of superimposed.," Ph.D. Thesis, Stanford Univ., Stanford, CA., Mar. 1985.

[31] I. Ziskind and M. Wax, "Maximum likelihood localization of multiple sources by alternating projection," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-36, no. 10, pp. 1553–1560, Oct. 1988.

[32] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-36, no. 4, pp. 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*, Feb. 1988, vol. 76, no. 2, pp. 203–205. Correction in *IEEE Proc.*, June 1988, vol. 76, no. 6, p. 734.

[34] J. A. Fessler, "Object-based 3-D reconstruction of arterial trees from a few projections,"*Ph.D. Thesis*,Stanford Univ.,Stanford,CA.,Aug. 1990.

[35] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. Ser. B*, vol. 48, no. 3, pp. 259–302, 1986.

[36] J. A. Cadzow, "Signal processing via least squares error modeling," *IEEE Signal Processing 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 Int. Conf. on Image Processing*, 1995, vol. 3, pp. 93–6.

[38] S. T. Jensen, S. Johansen, and S. L. Lauritzen, "Globally convergent algorithms for maximizing a likelihood function," *Biometrika*, vol. 78, no. 4, pp. 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. Comput. Assist. Tomogr.*, vol. 10, no. 5, pp. 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. Comput. Assist. Tomo.*, vol. 10, no. 2, pp. 296–301, Mar. 1986.

[41] G. T. Herman, *Image Reconstruction from Projections: The Fundamentals of Computerized Tomography*. New York: Academic, 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 Nucl. Sci. Symp. Med. Imag. Conf.*, 1995, vol. 2, pp. 1292–1296.

[43] J. A. Fessler, "Resolution properties of regularized image reconstruction methods," Commun. and Signal Processing Lab., Dept. EECS, Univ. of Michigan, Ann Arbor, MI, Tech. Rep. 297, pp. 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 Nucl. Sci. Symp. Med. Imag. Conf.*, 1994, vol. 4, pp. 1915–1919.

[45] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers, "On complete data spaces for PET reconstruction algorithms," *IEEE Trans. Nucl. Sci.*, vol. 40, no. 4, pp. 1055–1061, Aug. 1993.

[46] J. A. Fessler, "ASPIRE 3.0 user's guide: A sparse iterative reconstruction library," Commun. and Sign. Proc. Lab., Dept. EECS, Univ. of Michigan, Ann Arbor, MI, Tech. Rep. 293, pp. 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," Med. Image Processing. Group, Dept. Radiol., Univ. of Pennsylvania, Tech. Rep. MIPG119, Feb. 1987.

[49] J. A. O'Sullivan, "Roughness penalties on finite domains," *IEEE Trans. Image Processing*, Sept. 1995, vol. 4, no. 9, pp. 1258–68.