Alternating Minimization Algorithms for Transmission Tomography

Joseph A. O'Sullivan and Jasenka Benac*
Electronic Systems and Signals Research Laboratory
Department of Electrical and Systems Engineering
Washington University
St. Louis, MO 63130
jao@wustl.edu

May 30, 2006

Abstract

A family of alternating minimization algorithms for finding maximum likelihood estimates of attenuation functions in transmission x-ray tomography is described. The model from which the algorithms are derived includes polyenergetic photon spectra, background events, and nonideal point spread functions. The maximum likelihood image reconstruction problem is reformulated as a double minimization of the I-divergence. A novel application of the convex decomposition lemma results in an alternating minimization algorithm that monotonically decreases the objective function. Each step of the minimization is in closed form. The family of algorithms includes variations that use ordered subset techniques for increasing the speed of convergence. Simulations demonstrate the ability to correct the cupping artifact due to beam hardening and the ability to reduce streaking artifacts that arise from beam hardening and background events.
Key Words: transmission tomography, alternating minimization algorithms, maximum likelihood, image reconstruction, beam hardening

1 Introduction

Lange and Carson [20] cast the problem of image reconstruction in transmission tomography as a maximum likelihood estimation problem and derived an expectation-maximization algorithm to estimate the most likely image. Since their fundamental contribution, other researchers have introduced improvements over the years (for a review, see Fessler [16]). These include improvements in the approximate maximization step (M-step), increasing the speed of convergence (for example, using ordered subsets), and improving image quality through the use of regularization techniques.

This paper has several contributions: a reformulation of the maximum likelihood estimation problem as a double minimization of an I-divergence; a novel application of the convex decomposition lemma that yields an exact M-step; a family of algorithms that has been implemented in a range of medical and industrial applications; reduced complexity versions of the algorithm including ordered subsets and a version with fewer backprojections; and convergence analysis of the algorithm.

The reformulation of the maximum likelihood estimation problem starts with the equivalence between maximizing log-likelihood and minimizing a closely related I-divergence, which can be viewed as the generalization of relative entropy to arbitrary positive-valued distributions. This equivalence has been noted in the literature and relies on a Poisson data model. The reformulation continues by rewriting the I-divergence as the

*This work was funded in part by the National Cancer Institute of the National Institutes of Health under research grant R01CA75371 (J. F. Williamson, P. I.).

result of minimizing an I-divergence over a larger family of distributions. This second step of lifting the problem to a higher dimensional space is equivalent to the corresponding step in the expectation-maximization (EM) algorithm. Our formulation in terms of I-divergence allows us to connect this approach to a more general class of optimization problems, namely those that minimize I-divergence over a linear family in the first variable and over an exponential family in the second variable. Alternating minimization algorithms that arise in optimization have been studied extensively [7, 28, 30, 31, 4] and have been applied to medical imaging problems, especially to emission and transmission tomography [20, 21, 3, 22, 29, 38].

The formulation of Lange and Carson [20], is equivalent to a double minimization, one corresponding to the E-step and one to the M-step of the EM algorithm, which is itself a special case of alternating minimization algorithms. However, their M-step did not yield closed form updates. In order to overcome this, we develop a novel application of the convex decomposition lemma to the minimization over the exponential family. This yields an image update (M-step) that is closely related to the generalized iterative scaling algorithm of Darroch and Ratcliff [8]. A form of this lemma was described by De Pierro [10, 11] and was used by Lange and Fessler [21] to derive their convex algorithm.

We apply our approach to a polyenergetic model for X-ray CT data, in the presence of background events, to obtain a new family of image reconstruction algorithms. X-ray CT is one of the most widely used medical imaging modalities; it is used in industrial settings, including for nondestructive testing of airplane parts and for imaging of gas-solid risers in chemical engineering; and it is used in security applications, especially for baggage inspection. Metal and other high density attenuators are often present in the field of view; in patient imaging these can be known objects such as radiation brachytherapy applicators, orthopedic implants, or dental implants. In emergency applications, such attenuators may be unknown. In medical applications, the use of microCT for small animal imaging is increasingly important. Many of these applications require quantitative results, motivating improved data models and reconstruction algorithms. The algorithm described in this paper has led to the derivation of new algorithms: for problems with missing data, including region of interest tomography and tomography problems with missing projections [33]; for problems with known objects as described above [24, 25]; for the industrial applications mentioned; and for multiple energy data collection scenarios [2, 32].

We describe a Poisson model for X-ray CT measured data in Section 2 that incorporates polyenergetic attenuation and background events. The maximum likelihood estimation problem is reformulated as a double minimization I-divergence problem in Section 3. An alternating minimization algorithm is derived from this double minimization in Section 4 after the introduction of the convex decomposition lemma. Convergence properties (including global monotonic convergence, the proof of which is in the appendix), relationships to other algorithms, and a discussion of properties of the algorithm are included here as well. For example, we show that fixed points of the algorithm satisfy the first order Kuhn-Tucker conditions. After addressing the limitations of the model, numerical simulations are used to demonstrate key properties of the algorithm, including the correction of the cupping artifact due to beam hardening and the ability to reconstruct multiple components with dual energy data collection scenarios.

2 Polyenergetic Image Reconstruction Problem

Our algorithm derivation is based on a statistical model for the available data. This model accommodates the polyenergetic nature of X-ray beams, the existence of background events, Beer's law, and a realistic model for the known point spread function. The underlying continuous problem is discretized to yield the assumed model, with integrals and operators discretized appropriately. Within the limitations of the model (addressed in Section 5), the image reconstruction problem is formulated as an optimization (maximum likelihood) problem in statistical estimation theory. This model and maximum likelihood formulation are similar to that used by De Man, et al. [9].

Denote the transmission data by $d(y)$, where the variable y indexes source-detector pairs. The individual detector readings are assumed to be independent Poisson random variables with means $g(y : \mu)$, where we used the colon notation [45] to indicate the function is parameterized by the attenuation function, μ. The

Poisson log-likelihood function is

$$l(\mathbf{d} : \mu) \quad = \quad \sum_{y \in \mathcal{Y}} \left[d(y) \ln g(y : \mu) - g(y : \mu) \right]. \tag{1}$$

The data means are modeled as

$$g(y : \mu) = \sum_{E} I_0(y, E) \exp \left[- \sum_{x \in \mathcal{X}} h(y|x) \mu(x, E) \right] + \beta(y), \tag{2}$$

where the outer sum is over discrete energies of the X-ray photons. The summation in the exponent represents the forward projection of the attenuation function. The point spread function, $h(y|x)$, (in mm) accounts for effects such as finite detector size, the model for the discretization of the attenuation function (including pixel or voxel size and shape), the divergence of the X-ray beam, and the source-detector geometry.

The attenuation function $\mu(x, E)$ (in mm^{-1}) is indexed by image space coordinates, x, and by X-ray photon energy, E (nominally with units of keV). We envision a small number, I, of different types of materials indexed by i,

$$\mu(x, E) = \sum_{i=1}^{I} \mu_i(E) c_i(x), \tag{3}$$

with known linear attenuation coefficients $\mu_i(E)$ in mm^{-1} and relative partial densities $c_i(x)$ [41]. For pure linear combinations, the relative partial densities are nonnegative and sum to one. Our model restricts the values of $c_i(x)$ to be nonnegative, but does not enforce a sum constraint in order to allow the $\mu_i(E)$ to merely span the set of allowable attenuation functions $\mu(x, E)$. Our model for $\mu(x, E)$ (3) is equivalent to having terms $(\mu/\rho)_i(E)\rho_i(x)$, where $(\mu/\rho)_i(E)$ is the mass attenuation coefficient (usually given in cm^2/g and $\rho_i(x)$ is the partial density (in g/cm^3, with $h(y|x)$ in cm) of the ith constituent. The model (3) is related to others in the literature [9, 12, 13, 37, 43].

The function $\beta(y)$ in (2) denotes the mean number of background events and is assumed to be nonnegative and known; $\beta(y)$ includes scattered photons that contribute to detector readings arising from primary photon collisions either along the source-detector ray path or elsewhere in the patient volume (or accidental coincidences in transmission scans used in emission computed tomography [26, 14]). $I_0(y, E)$ is the mean number of events for source-detector pair y at energy E, in the absence of attenuation and background events.

Maximum likelihood estimation problem statement: *Find* $\{c_i(x),\ i = 1, 2, \ldots, I\ , x \in \mathcal{X}\}$ *that maximize the log-likelihood function (1) using (2) and (3), subject to nonnegativity of all* $c_i(x)$.

For typical scanner detector arrays and image resolution specifications, this estimation problem as stated is ill-posed. There are many standard techniques used to regularize such problems, including the use of constraints, roughness penalty functions, prior likelihood functions, sieves, stopping criteria, and complexity penalty functions (see Fessler [16] or O'Sullivan, Blahut, and Snyder [29] for a survey of standard regularization techniques). Each of these techniques has its own merits. Natural constraints may include that a given voxel has only one constituent (for each x, only one $c_i(x)$ is nonzero) and that neighboring voxels tend to have the same constituent. Roughness penalties are used extensively in transmission tomography [11, 14, 15, 21], and have been described in rather general settings as well [27].

Since our purpose is to introduce the family of alternating minimization algorithms and to identify their basic properties, no regularization techniques are used. However, it is straightforward to modify these algorithms to include penalties using standard techniques. Two possibilities are graph coloring methods accounting for induced neighborhood structure [27] and convex decompositions as used by Erdogan and Fessler [14, 15] and Lange and Fessler [21] based in part on the work of De Pierro [10, 11]. The main contributions of this paper are in the novel reformulation and decomposition of the log-likelihood functional and the resulting family of iterative algorithms.

The algorithms in this paper lead directly to fast implementations using ordered subset techniques [17, 18]. While we have implemented and extensively tested computer code using ordered subset techniques, all images shown here use the full iterations as described.

3 Reformulation of Problem Statement

The derivation of the algorithms is based on a mixture of novel and classic ideas. The novel component of our approach is derivation of iterative algorithms by minimizing an I-divergence [28, 30]. The classic ideas are from Darroch and Ratcliff [8], who studied iterative algorithms for minimizing relative entropy (maximizing entropy) subject to equality constraints. Our algorithm for transmission tomography is an extended version of Darroch and Ratcliff's generalized iterative scaling algorithm. Our algorithm is based on a transformation of the maximum likelihood estimation problem into a double minimization of an I-divergence between two functions. The first function is constrained to have values in a linear family while the second function is constrained to have values in an exponential family. The iterative algorithm alternates between updating the two functions.

The terms in the log-likelihood function (1) that depend on $\mathbf{c}$ (the vector of functions $c_i(x)$ in (3)) are the negative of the corresponding terms in the I-divergence

$$I(\mathbf{d}\|\mathbf{g}(y:\mathbf{c})) = \sum_{y\in\mathcal{Y}} \left(d(y) \ln \frac{d(y)}{g(y:\mathbf{c})} - d(y) + g(y:\mathbf{c}) \right). \tag{4}$$

Thus, minimizing the I-divergence (4) over $\mathbf{c}$ is equivalent to maximizing the log-likelihood function (1) over $\mathbf{c}$.

Exponential families can be described as log-linear and in some cases log-convex; a convex sum of the logarithms of two elements in such a family equal the logarithm of another element of the family. For our problem, we consider the following exponential family whose elements are parameterized by $\{c_i(x),\ i = 1, 2, \ldots, I\}$. This definition includes a dummy energy value, $E = 0$, to account for the background events.

Definition 3.1 *The* **exponential family** $\mathcal{E}$ *is the set*

$$\mathcal{E} = \{\mathbf{q} \quad : \quad q(y,E) = I_0(y,E) \exp\left(-\sum_{x\in\mathcal{X}} h(y|x) \sum_{i=1}^{I} \mu_i(E) c_i(x) \right),\ \textit{for } E \neq 0, \tag{5}$$

$$q(y,0) = \beta(y)\}\,. \tag{6}$$

The exponential family defines the model used for the data; $q(y,E)$ is the mean number of counts for source-detector pair y that have energy E. The mean number of total counts for y, as already used in the I-divergence (4), is $g(y:\mathbf{c})$ where

$$\begin{aligned} \mathbf{g}(y:\mathbf{c}) &= \sum_{E} q(y,E) \\ &= \sum_{E\neq 0} I_0(y,E) \exp\left(-\sum_{i=1}^{I} \sum_{x\in\mathcal{X}} h(y|x) \mu_i(E) c_i(x) \right) + \beta(y). \end{aligned} \tag{7}$$

Below, all summations over E are assumed to include the dummy variable $E = 0$ as well, that term corresponding to the background events. Corresponding to $E = 0$, we set $\mu_i(0) = 0$ for all i. While general linear families are addressed by O'Sullivan [30], here we restrict attention to a simplified linear family that is specified as those nonnegative functions whose marginals on y equal $d(y)$.

Definition 3.2 *The* **linear family** *whose marginals equal* $\mathbf{d}$ *is*

$$\mathcal{L}(\mathbf{d}) = \{p(y,E) \geq 0 : \sum_{E} p(y,E) = d(y)\}. \tag{8}$$

One view is that $p(y,E)$ is the mean number of counts for source-detector pair y that have energy E. The set $\mathcal{L}$ imposes the constraint that the total mean number of counts for y equals the measurement $d(y)$. The counts with energy E for source-detector pair y are not directly measured and correspond to hidden data used in the EM algorithm. The algorithm produces an estimate of $p(y,E)$ at each iteration.

Lemma 1 *The I-divergence (4) may be written in the variational form*

$$I(\mathbf{d}\|\mathbf{g}(y:\mathbf{c})) = \min_{\mathbf{p}\in\mathcal{L}(\mathbf{d})} I(\mathbf{p}\|\mathbf{q}), \tag{9}$$

where

$$I(\mathbf{p}\|\mathbf{q}) = \sum_y \sum_E \left(p(y,E) \ln \frac{p(y,E)}{q(y,E)} - p(y,E) + q(y,E) \right). \tag{10}$$

Proof: Introduce a Lagrange multiplier $\lambda(y)$ to enforce the equality in the definition of $\mathcal{L}(\mathbf{d})$ in (8) to get the Lagrangian, L,

$$L = \sum_y \sum_E \left(p(y,E) \ln \frac{p(y,E)}{q(y,E)} - p(y,E) + q(y,E) \right) + \sum_{y\in\mathcal{Y}} \lambda(y) \left(\sum_E p(y,E) - d(y) \right). \tag{11}$$

Minimizing over $p(y,E)$ and solving for $\lambda(y)$ to enforce equality yields $p(y,E) = 0$ if $q(y,E) = 0$ (defining $I(0||0) = 0$) and if $q(y,E) \neq 0$,

$$p(y,E) = d(y) \frac{q(y,E)}{\sum_{E'} q(y,E')}. \tag{12}$$

Substituting this back into the I-divergence yields the equality in the Lemma statement. Nonnegativity of $p(y,E)$ is inherent. That completes the proof.

Using this form, the maximum likelihood estimation problem may be rewritten as

$$\min_{\mathbf{q}\in\mathcal{E}} \min_{\mathbf{p}\in\mathcal{L}(\mathbf{d})} I(\mathbf{p}\|\mathbf{q}), \tag{13}$$

subject to the inequality constraints $c_i(x) \geq 0$, for all (i,x). This double minimization leads to an alternating minimization algorithm [7, 28, 31], where the iterations alternate between estimating $\mathbf{p} \in \mathcal{L}(\mathbf{d})$ and $\mathbf{q} \in \mathcal{E}$. Any of the class of algorithms described below based on iterating between these two minimizations is referred to as an alternating minimization algorithm.

There is a general class of algorithms that consists of a double minimization as in (13), where the first variable is minimized over a linear family and the second over an exponential family. This class includes the expectation maximization algorithm for emission tomography. The linear family can be updated analytically if it corresponds to a constraint on the marginals of a function, as is the case for both the emission and transmission tomography problems. Otherwise, the minimization over the linear family may require subiterations. Similarly, the exponential family can be updated analytically if it corresponds to the product of components, as is the case for emission tomography but not for transmission tomography. Thus for transmission tomography, the exponential family minimizations require analytical subiterations. The primary algorithm recommended below uses one such subiteration for each iteration of the linear family.

It is important to recognize that the subiterations for $\mathbf{q}$ are themselves iterations of another alternating minimization algorithm with analytical updates; while one subiteration is recommended below, any number of these subiterations may be run for each iteration for $\mathbf{p}$, thereby avoiding some computations. This algorithm is an extension of the generalized iterative scaling algorithm of Darroch and Ratcliff [8], who studied the problem of maximizing entropy subject to linearity constraints. They demonstrated the equivalence of that problem to a maximum likelihood estimation problem within an exponential family. Maximizing entropy was formulated as minimizing relative entropy, relative entropy being equivalent to I-divergence for probability distributions.

For our problem, given an estimate $\hat{\mathbf{p}}$, the first order necessary (Kuhn-Tucker) conditions for $\mathbf{q} \in \mathcal{E}$ to be a minimizer are that for each (x,i) such that $c_i(x) > 0$

$$\sum_{y\in Y} \sum_E \hat{p}(y,E) h(y|x) \mu_i(E) = \sum_{y\in Y} \sum_E q(y,E) h(y|x) \mu_i(E); \tag{14}$$

and for each (x, i) such that $c_i(x) = 0$,

$$\sum_{y \in Y} \sum_E \hat{p}(y, E) h(y|x) \mu_i(E) \geq \sum_{y \in Y} \sum_E q(y, E) h(y|x) \mu_i(E). \tag{15}$$

The summations in these equations are called backprojections; note that they refer to backprojections of the measured or estimated flux, not the usual attenuation sinogram. The backprojection here includes the usual spatial backprojection (sum over $y \in \mathcal{Y}$ after multiplication by $h(y|x)$) and a backprojection over energies to constituents (sum over E after multiplication by $\mu_i(E)$). Equations (14) and (15) imply that the backprojected member of the linear family must equal the backprojected member of the exponential family over the set of (x, i) such that $c_i(x) > 0$.

The following lemma extends the correspondence used by Darroch and Ratcliff [8] to derive their algorithm.

Lemma 2 *Fix a nonnegative* $\mathbf{p}$. *The following two problems are equivalent:*
(i) minimize $I(\mathbf{p}\|\mathbf{q})$ *over all* $\mathbf{q} \in \mathcal{E}$*; and*
(ii)minimize $I(\mathbf{q}\|I_0)$ *over all nonnegative* $\mathbf{q}$ *subject to the constraints*

$$\sum_{y \in Y} \sum_E p(y, E) h(y|x) \mu_i(E) = \sum_{y \in Y} \sum_E q(y, E) h(y|x) \mu_i(E) \tag{16}$$

for all (x, i).

Note that in this lemma statement, there are no nonnegativity constraints on $c_i(x)$. For probability distributions, the second problem is a standard maximum entropy or minimum relative entropy problem subject to linearity constraints. The first problem is a minimum relative entropy over an exponential family. The minimum relative entropy solution subject to the linearity constraint is a member of the exponential family that minimizes the relative entropy to $\mathbf{p}$. This lemma is one expression of the duality of these two problems.

The presence of the nonnegativity constraints $c_i(x) \geq 0$ changes Lemma 2: minimizing $I(\mathbf{p}\|\mathbf{q})$ over all $\mathbf{q} \in \mathcal{E}$ subject to $c_i(x) \geq 0$ is equivalent to minimizing $I(\mathbf{q}\|I_0)$ over all nonnegative $\mathbf{q}$ subject to the Kuhn-Tucker conditions (14) and (15) with $\hat{p}$ being replaced by p.

4 Derivation of Iterative Algorithm

Let $\mathcal{P} = \{\mathbf{r} : r_i \geq 0, \sum_i r_i = 1\}$. We refer to the following lemma as the convex decomposition lemma.

Lemma 3 *Suppose that* f *is a convex function defined on a convex cone* $\mathcal{D} \subset \mathbf{R}^n$. *Given* $\mathbf{x}_i \in \mathcal{D}, i = 1, 2, \ldots,$

$$f\left(\sum_i \mathbf{x}_i\right) \leq \sum_i r_i f\left(\frac{1}{r_i}\mathbf{x}_i\right), \tag{17}$$

for all $\mathbf{r} \in \mathcal{P}$, *with* $r_i > 0$ *for all* i. *If* f *is strictly convex, equality holds if and only if* $\frac{1}{r_i}\mathbf{x}_i = \mathbf{x}$ *is independent of* i.

The proof follows from Jensen's inequality applied to $f\left(\sum_i \frac{r_i}{r_i}\mathbf{x}_i\right)$. This lemma is used by Lange and Fessler [21] and by Erdogan and Fessler [15], based on work of De Pierro [10, 11], for deriving algorithms in transmission tomography. The lemma can be viewed as the basis for the expectation-maximization algorithm and can be applied in a variety of ways [30]. We apply the convex decomposition lemma in a new way.

Note that the terms involving $c_i(x)$ from $I(\hat{\mathbf{p}}\|\mathbf{q})$ are

$$\sum_{y\in Y}\sum_{E}\hat{p}(y,E)\sum_{x}\sum_{i}h(y|x)\mu_i(E)c_i(x)+\sum_{y\in Y}\sum_{E}I_0(y,E)\exp\left(-\sum_{i}\sum_{x}h(y|x)\mu_i(E)c_i(x)\right) =$$

$$\sum_{y\in Y}\sum_{E}\hat{p}(y,E)\sum_{x}\sum_{i}h(y|x)\mu_i(E)c_i(x)+\sum_{y\in Y}\sum_{E}\hat{q}(y,E)\exp\left(\sum_{i}\sum_{x}h(y|x)\mu_i(E)[\hat{c}_i(x)-c_i(x)]\right) \tag{18}$$

for any estimate of $\mathbf{q}$ denoted $\hat{\mathbf{q}}$ with corresponding estimates $\hat{c}_i(x)$. As has been noted by Lange and Fessler [21], the function

$$f(y,E,t)=t\hat{p}(y,E)+\hat{q}(y,E)\exp(-t) \tag{19}$$

is convex in the variable t. The convex decomposition lemma yields

$$\sum_{y\in Y}\sum_{E}\hat{p}(y,E)\sum_{x}\sum_{i}h(y|x)\mu_i(E)c_i(x)+\sum_{y\in Y}\sum_{E}I_0(y,E)\exp\left(-\sum_{i}\sum_{x}h(y|x)\mu_i(E)c_i(x)\right) \leq$$

$$\sum_{y\in Y}\sum_{E}\sum_{i}\sum_{x}\left[\hat{p}(y,E)h(y|x)\mu_i(E)c_i(x)+r(x,i|y,E)\hat{q}(y,E)\exp\left(\frac{h(y|x)\mu_i(E)}{r(x,i|y,E)}[\hat{c}_i(x)-c_i(x)]\right)\right] \tag{20}$$

for all $r(x,i|y,E)>0$ such that

$$\sum_{x}\sum_{i}r(x,i|y,E)\leq 1. \tag{21}$$

Note the inequality in (21); this minor extension of the convex decomposition lemma is valid due to the possibility of adding a dummy x variable (again denoted 0) such that $\hat{c}_i(0)-c_i(0)=0$ for each i. Equality is achieved in (20) if

$$\frac{h(y|x)\mu_i(E)}{r(x,i|y,E)}[\hat{c}_i(x)-c_i(x)] \tag{22}$$

is only a function of (y,E). One clear possibility for this is if the algorithm converges and $\hat{c}_i(x)=c_i(x)$.

To derive an alternating minimization algorithm for X-ray transmission CT, set

$$r(x,i|y,E)=\frac{h(y|x)\mu_i(E)}{Z_i(x)}, \tag{23}$$

where $Z_i(x)$ are chosen to enforce the constraint (21). We have found that some scanner geometries allow a selection of $Z(x)$ in the monoenergetic version of this algorithm to achieve equality in (21) for all y. In general, the $Z_i(x)$ must be large enough, one such choice being

$$Z_i(x)=Z_0=\max_{(y,E)}\sum_{x\in\mathcal{X}}\sum_{i}\mu_i(E)h(y|x). \tag{24}$$

The resulting decoupled objective function is

$$\sum_{i}\sum_{x}\sum_{y\in Y}\sum_{E}\left[\hat{p}(y,E)h(y|x)\mu_i(E)c_i(x)+\hat{q}(y,E)h(y|x)\mu_i(E)\frac{1}{Z_i(x)}\exp\left(Z_i(x)[\hat{c}_i(x)-c_i(x)]\right)\right], \tag{25}$$

which can be minimized in closed form over $c_i(x)$. This motivates the algorithm. At iteration k, assume that $\hat{\mathbf{q}}^{(k)}$ is given; then $\hat{\mathbf{p}}^{(k)}$ is determined by a direct minimization as in (12). The objective function (25) uses $\hat{\mathbf{p}}^{(k)}$ and $\hat{\mathbf{q}}^{(k)}$ (with the corresponding $\hat{c}_i^{(k)}(x)$). The direct minimization then yields $\hat{c}_i^{(k+1)}(x)$ and thus $\hat{\mathbf{q}}^{(k+1)}$.

4.1 Alternating Minimization Iterations

Set $k = 0$. Select an initial condition for each $\hat{c}_i^{(0)}(x)$. Compute the current estimate for the function $\mathbf{q}$ in the exponential family $\mathcal{E}$

$$\hat{q}^{(k)}(y, E) = I_0(y, E) \exp\left[- \sum_{x \in \mathcal{X}} \sum_i \mu_i(E) h(y|x) \hat{c}_i^{(k)}(x) \right]. \tag{26}$$

Compute the current estimate of the function $\mathbf{p}$ in the linear family $\mathcal{L}$

$$\hat{p}^{(k)}(y, E) = \hat{q}^{(k)}(y, E) \frac{d(y)}{\sum_{E'} \hat{q}^{(k)}(y, E')}. \tag{27}$$

Compute the backprojection of the current estimates of $\mathbf{p}$ and $\mathbf{q}$

$$\tilde{b}_i^{(k)}(x) = \sum_{y \in \mathcal{Y}} \sum_E \mu_i(E) h(y|x) \hat{p}^{(k)}(y, E) \tag{28}$$

$$\hat{b}_i^{(k)}(x) = \sum_{y \in \mathcal{Y}} \sum_E \mu_i(E) h(y|x) \hat{q}^{(k)}(y, E). \tag{29}$$

Update the estimate of the relative partial densities (by minimizing (25))

$$\hat{c}_i^{(k+1)}(x) = \hat{c}_i^{(k)}(x) - \frac{1}{Z_i(x)} \ln \left(\frac{\tilde{b}_i^{(k)}(x)}{\hat{b}_i^{(k)}(x)} \right). \tag{30}$$

If $\hat{c}_i^{(k+1)}(x) < 0$, set $\hat{c}_i^{(k+1)}(x) = 0$.
Set $k = k + 1$. Check for convergence and iterate if necessary.

4.2 Discussion

- The computational complexity of the algorithm is proportional to the number of constituents $c_i(x)$. For each constituent i, one forward projection and two backprojections are required.

- As noted briefly above Lemma 2, the iterations for $\mathbf{q}$, computing $\hat{c}_i^{(k+1)}(x)$, may be viewed as subiterations, with one such subiteration performed per iteration on $\mathbf{p}$. With multiple such subiterations per iteration of $\mathbf{p}$, the backprojections of $\hat{\mathbf{p}}^{(k)}$ are avoided, reducing the computational complexity due to the backprojections by a factor of 2 per iteration. Note that the monotonicity of the I-divergence is maintained for each subiteration of this reduced complexity algorithm.

- The estimate for $q(y, 0)$ is always $\beta(y)$. Using this, the iterations for $\hat{p}^{(k)}(y, E)$ may be rewritten as

$$\hat{p}^{(k)}(y, E) = \hat{q}^{(k)}(y, E) \frac{d(y)}{\sum_{E' \neq 0} \hat{q}^{(k)}(y, E') + \beta(y)}. \tag{31}$$

  The function $\hat{p}^{(k)}(y, E)$ equals the expected number of counts from $d(y)$ attributed to uncollided photons of energy E and not due to background, given the current estimate of the $c_i(x)$. While it is not directly used in the algorithm,

$$\hat{p}^{(k)}(y, 0) = \beta(y) \frac{d(y)}{\sum_{E' \neq 0} \hat{q}^{(k)}(y, E') + \beta(y)} \tag{32}$$

  is the expected number of counts due to the background. These are analogous to conditional means computed by Ollinger [26, p. 94].

- Recall that $\mu_i(0) = 0$, so the summations in the backprojections to compute $\tilde{b}_i^{(k)}(x)$ and $\hat{b}_i^{(k)}(x)$ are only over $E \neq 0$.

- If the monoenergetic model of the X-ray source is assumed (*i.e.* there is only one energy level, E), the algorithm above reduces to estimating only one image, $c(x)$. In this case, $\tilde{b}(x)$ becomes the backprojection of the measured data, $d(y)$, which can be precomputed so the iterations require just one forward projection and one backprojection. Although this model does not help in reducing the beamhardening artifacts, it still may be of interest because it accounts for the randomness in the measure data and thus may mitigate noise artifacts for relatively low computational cost.

- The hard threshold on $\hat{c}_i^{(k+1)}(x)$ is an enforcement of the nonnegativity constraint at every iteration. That is, the minimization of (25) is a constrained minimization.

- The algorithm can be modified easily to include known values for $c_i(x)$ in any region. In the update for the $c_i(x)$, the known values are not updated. This is an extension of the method described by Wang, et al. [38] for iterative deblurring in the presence of high density attenuators with known attenuation maps. Our results using this approach have been extremely encouraging; see Williamson, et al. [41].

- When the incident energy spectrum, $I_0(y, E)$, is constant for all source-detector pairs, y, we were able to estimate only one constituent without imposing strong constraints such as $c_1(x) + c_2(x) = 1$. We were able to estimate two constituents when the incident energy spectrum varies with y as is the case in dual-energy CT scanning where two scans are obtained using different X-ray tube voltage settings (see Section 6). Stonestrom, *et al.* [36] present a strong argument based on the underlying physics and experimental data for using two constituents. We hypothesize that I components can be estimated if I scans are obtained using I different X-ray tube voltage settings.

- Using the value Z_0 specified above may lead to slow convergence, so our simulations often use smaller values to accelerate the convergence. Changing the value $Z_i(x)$ with the iteration number is possible; decreasing values of $Z_i(x)$ may lead to improved convergence rate, while increasing values may lead to slower convergence rate.

- The algorithm may be extended to estimate parameters in the mean number of background events, $\beta(y)$. For example, if $\beta(y)$ is a constant, estimating its value is a straightforward extension.

4.3 Convergence Analysis

We look at several important properties related to the convergence of the algorithm.

Theorem 4 *The algorithm presented in Subsection 4.1 monotonically increases the log-likelihood function, or equivalently, decreases the I-divergence*

$$I(\mathbf{d}\|\mathbf{g}(y : \hat{\mathbf{c}}^{(k)})) \geq I(\mathbf{d}\|\mathbf{g}(y : \hat{\mathbf{c}}^{(k+1)})) \tag{33}$$

with equality if and only if the Kuhn-Tucker conditions (14) and (15) are satisfied.

The functions $\tilde{b}_i^{(k)}(x)$ and $\hat{b}_i^{(k)}(x)$, defined in Subsection 4.1, are the backprojections of the current estimates for $\mathbf{p}$ and $\mathbf{q}$, respectively. This theorem states a fundamental property of minimizing I-divergence in the second variable over an exponential family, namely that the moments of the estimate should match the moments of the first variable; the moments are defined in terms of the linear function in the exponent of the exponential family. The proof is in Appendix A.

Note that several properties of the algorithm immediately follow from the proof of Theorem 4 in the appendix. The function $\tilde{Z}_i^{(k)}(x)$ is defined in the appendix.

Theorem 5 *Assume that $I(\mathbf{d}\|g(y:\hat{\mathbf{c}}^{(1)}))$ is finite. Then*
(a) The sums

$$\sum_{k=1}^{\infty} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{p}}^{(k+1)}) \tag{34}$$

and

$$\sum_{k=1}^{\infty} I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\|\frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\right) \tag{35}$$

are finite.
(b) The successive I-divergences $I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{p}}^{(k+1)})$ converge to zero.
(c) The successive I-divergences $I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\|\frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\right)$ converge to zero.
(d) The set of limit points of the iterates $c_i^{(k)}(x)$ is a connected set.

Part (a) of this theorem follows from summing the inequality in (53) to get

$$\begin{aligned} I(\mathbf{d}\|g(y:\hat{\mathbf{c}}^{(1)})) - I(\mathbf{d}\|g(y:\hat{\mathbf{c}}^{(K+1)})) &= \\ \sum_{k=1}^{K} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}) - I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) &+ \\ \sum_{k=1}^{K} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) - I(\hat{\mathbf{p}}^{(k+1)}\|\hat{\mathbf{q}}^{(k+1)}) &\geq \qquad (36) \\ \sum_{k=1}^{K} I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\|\frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\right) + \sum_{k=1}^{K} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{p}}^{(k+1)}) & . \qquad (37) \end{aligned}$$

A finite value of $I(\mathbf{d}\|g(y:\hat{\mathbf{c}}^{(1)}))$ implies a finite value of the sums. This in turn implies that each term in the sums converges to zero. Connectedness of the limit set of the iterations follows from convergence of $I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\|\frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}\right)$ to zero which implies that the differences between successive iterates of $\hat{c}_i^{(k)}(x)$ converge to zero. This completes the proof of the theorem.

Theorem 6 *(a) Let $\Gamma^* = \{\gamma_i^*(x), i = 1, 2, \ldots, I\}$ be any set that satisfies the Kuhn-Tucker conditions (14) and (15). Then Γ^* is a fixed point of the iterations.*
(b) Let $\mathbf{c} = \{c_i^*(x), i = 1, 2, \ldots, I\}$ be any set that is a fixed point of the alternating minimization iterations. Then $\mathbf{c}$ satisfies the Kuhn-Tucker conditions (14) and (15).

Part (a) follows from the modified update expression derived in (47) in the appendix

$$\hat{c}_i^{(k+1)}(x) = \hat{c}_i^{(k)}(x) - \frac{1}{\tilde{Z}_i^{(k)}(x)} \ln \frac{\tilde{b}_i^{(k)}(x)}{\hat{b}_i^{(k)}(x)}. \tag{38}$$

Let $\hat{c}_i^{(k)}(x) = \gamma_i^*(x), i = 1, 2, \ldots, I$, where Γ^* satisfies the Kuhn-Tucker conditions. For the $x \in \mathcal{X}$ such that $\gamma_i^*(x) > 0$, $\tilde{b}_i^{(k)}(x) = \hat{b}_i^{(k)}(x)$, so the update is zero. For the $x \in \mathcal{X}$ such that $\hat{c}_i^{(k)}(x) = \gamma_i^*(x) = 0$, the values of $\tilde{Z}_i^{(k)}(x)$ are defined so that $\hat{c}_i^{(k+1)}(x) = 0$, and thus we have a fixed point.

For part (b) of the theorem, let $\hat{c}_i^{(k)}(x) = c_i^*(x), i = 1, 2, \ldots, I$ be a fixed point. Then, for $x \in \mathcal{X}$ such that $c_i^*(x) > 0$, $\tilde{b}_i^{(k)}(x) = \hat{b}_i^{(k)}(x)$ because $\hat{c}_i^{(k+1)}(x) = \hat{c}_i^{(k)}(x)$ and $\tilde{Z}_i^{(k)}(x) > 0$ and finite by definition. For $x \in \mathcal{X}$ such that $c_i^*(x) = 0$, $c_i^*(x)$ being a fixed point immediately implies that $\tilde{b}_i^{(k)}(x) \geq \hat{b}_i^{(k)}(x)$. This proves the theorem.

These theorems provide the basis for understanding convergence properties of the algorithm. The set of limit points of $\hat{\mathbf{q}}^{(k)}$ form a connected set. Furthermore, $\hat{q}^{(k)}(y,E) \leq I_0(y,E)$ for all (y,E) so $\hat{\mathbf{q}}^{(k)}$ takes values in a compact set and the sequence has a convergent subsequence. If the minimum value of $d(y)$ is strictly positive, the minimum value of $I_0(y,E)$ is strictly positive, and for every (i,x) the term $\mu_i(E)h(y|x)$ is strictly positive for some (y,E), then all values of $c_i(x)$ are bounded above, and thus the iterations for $\{c_i(x), i = 1,2,\ldots,I\}$ remain in a compact set. If in addition the maximum over (y,E) of $\mu_i(E)h(y|x)$ for all (i,x) is finite, then the values of the limit points of $\hat{q}^{(k)}(y,E)$ are bounded below, as are the values of $\tilde{b}_i^{(k)}(x)$ and $\hat{b}_i^{(k)}(x)$; the values of $\tilde{b}_i^{(k)}(x)$ and $\hat{b}_i^{(k)}(x)$ are also then bounded above. Then $\ln\left(\tilde{b}_i^{(k)}(x)/\hat{b}_i^{(k)}(x)\right)$ is finite, and when viewed as a function of $c_i(x)$ is continuous; for any convergent subsequence, the corresponding values of $\ln\left(\tilde{b}_i^{(k)}(x)/\hat{b}_i^{(k)}(x)\right)$ converge. By property (c) of Theorem 5, if $\tilde{b}_i^{(k)}(x)$ remains bounded and strictly positive, then $\left(1/\tilde{Z}_i^{(k)}(x)\right)\ln\left(\tilde{b}_i^{(k)}(x)/\hat{b}_i^{(k)}(x)\right)$ converges to zero.

4.4 Ordered Subsets Alternating Minimization Algorithms

The convergence speed of this algorithm may be increased using ordered subset techniques [17], similar to those described by Kamphuis and Beekman [18] and Erdogan and Fessler [15]. The idea is to partition each iteration into subiterations, each of which uses only a subset of the measured data, $\{d(y) : y \in \mathcal{Y}_j\}$, to perform the update. The subsets form a partition of the data into disjoint subsets so that $\mathcal{Y}_j \cap \mathcal{Y}_l = \emptyset$ for all $j \neq l$ and $\cup_{j=1}^{J}\mathcal{Y}_j = \mathcal{Y}$. Each subiteration requires one partial forward projection and two partial backprojections per constituent, defined by the appropriate subset. The total number of computations per iteration is approximately the same as the full algorithm in Subsection 4.1. Convergence is sped up approximately by the number of subsets J.

More specifically, in subiteration j, the forward projection in (26) is performed only for $y \in \mathcal{Y}_j$, as is the update of the current estimate for p in (27). The backprojections in (28) and (29) involve only sums over $y \in \mathcal{Y}_j$. All elements of $c_i(x)$ are updated, but the normalizing function may depend on iteration number, $Z_{ij}(x)$. The $Z_{ij}(x)$ must satisfy a modified condition that

$$\sum_i \sum_x \frac{h(y|x)\mu_i(E)}{Z_{ij}(x)} \leq 1, \tag{39}$$

for all E and for all $y \in \mathcal{Y}_j$.

It is well known that ordered subsets algorithms are not guaranteed to converge, and that is the case here as well. This ordered subsets algorithm does increase convergence speed substantially. There are several strategies available to further improve convergence, three of which are mentioned here. One strategy is for the iterations to use a relaxation procedure to smooth between successive updates; see Ahn and Fessler [1]. A second strategy is to run a few iterations of the full algorithm (without ordered subsets) after getting close to convergence. A third strategy, and one that is not always possible, is to use a decreasing sequence of $Z_i^{(k)}(x)$ in place of $Z_i(x)$; this places increasing weights on the step-sizes with iteration number.

In order to focus on the new algorithm, the images shown below were computed using the full iterations from Subsection 4.1. We have implemented the ordered subset version with promising results [41].

4.5 Related Iterative Algorithms

Lange and Fessler [21] compare a gradient algorithm and a convex algorithm to the algorithms originally introduced by Lange and Carson. Lange and Fessler's gradient algorithm may be written as

$$\hat{c}^{(k+1)}(x) = \hat{c}^{(k)}(x)\frac{\hat{b}^{(k)}(x)}{b(x)}, \tag{40}$$

where $b(x)$ and $\hat{b}^{(k)}(x)$ are the backprojections of $d(y)$ and $\hat{q}^{(k)}(y)$ using $h(y|x)$, respectively. They refer to this as a gradient algorithm since it may further be rewritten as

$$\hat{c}^{(k+1)}(x) = \hat{c}^{(k)}(x) + \frac{\hat{c}^{(k)}(x)}{b(x)} \frac{\partial}{\partial c(x)} I(\mathbf{d} \| \mathbf{g}(y : \mathbf{c})). \tag{41}$$

Lange and Fessler's convex algorithm [21] is similar in spirit to but distinct from the alternating minimization algorithm introduced here. They address the monoenergetic case and apply the convex decomposition lemma directly to the negative log-likelihood function to obtain

$$\begin{aligned} -l(\mathbf{c}) &= \sum_y d(y) \sum_x h(y|x)c(x) + I_0(y) \exp\left(-\sum_x h(y|x)c(x)\right) \\ &\leq \sum_y \sum_x \left(d(y)h(y|x)c(x) + I_0(y)r(x|y) \exp\left(-\frac{h(y|x)c(x)}{r(x|y)}\right)\right), \end{aligned} \tag{42}$$

for any conditional probabilities $r(x|y)$. Given an estimate $\hat{c}^{(k)}$, the minimum over $r(x|y)$ is obtained as

$$\hat{r}^{(k)}(x|y) = \frac{h(y|x)\hat{c}^{(k)}(x)}{\sum_{x'} h(y|x')\hat{c}^{(k)}(x')}. \tag{43}$$

They then apply Newton's method, expanding $c(x)$ around the current estimate $\hat{c}^{(k)}(x')$. Their algorithm [21, Eq. (7)] requires one forward projection and two backprojections as well.

Nuyts et al. [23] describe a closely related algorithm for the monoenergetic case. Their algorithm essentially is

$$\hat{c}^{(k+1)}(x) = \hat{c}^{(k)}(x) - \frac{1}{Z}\left(\frac{b(x)}{\hat{b}^{(k)}(x)} - 1\right). \tag{44}$$

Since $\ln \alpha \leq \alpha - 1$, the convergence analysis in the appendix does not appear to apply to their algorithm. With Z large enough, there should not be a problem, but their algorithm appears to lack the global monotonicity of our algorithm.

Elbakri and Fessler [12, 13] describe an algorithm for the polyenergetic model. Their algorithm uses ordered subsets and a separable paraboloidal surrogate function. In addition, they explicitly recommend and implement a regularization method. As noted above, for the algorithm described here to be useful, regularization is required. The implementations described here use a single energy-dependent linear attenuation coefficient, reducing the need for regularization and highlighting the contributions of this paper.

5 Model Limitations

The algorithm presented in this paper is designed to solve a specific maximum likelihood estimation problem. As with all models, the model assumed is only approximate, capturing much but not all of the underlying physics. The actual data in transmission tomography are typically not Poisson distributed due to either system nonlinearities or fundamental physical considerations. For example, the solid state detectors currently used by modern CT scanners are energy-integrating detectors, not photon-counting detectors; for energy-integrating detectors a detected photon whose energy is E keV contributes E keV to the measurements. The mean of the measurements is accurately modeled using extensions of the equations here, but the log-likelihood function becomes a mixture of Poisson random variables. Some aspects of the underlying physics are discussed by Whiting [39, 40].

In our model, $h(y|x)$ is based on a discretization of the underlying line integrals. Typically the discretization assumes constant attenuation over a pixel or voxel and $h(y|x)$ includes an integral over the detector. In real measurements, both of these aspects are approximations. Actual detectors integrate photon flux at each

energy over the surface of the detector (not attenuation functions). Actual attenuation functions of interest are not constant over pixels or voxels.

The assumption that the mean number of background events, $\beta(y)$, is a known constant is rather simplistic. Scatter in transmission tomography depends in a complicated manner on the attenuation function, and hence is energy-dependent as well. More realistic models for scatter have been proposed, but to our knowledge no computationally attractive algorithms have been proposed based on these more complicated models. Our algorithms are easily extended to estimating constant values of β (independent of y).

6 Simulation Results

We present the results of simulations in this section, demonstrating some key features of the proposed algorithm, namely its ability to account for polyenergetic models and for background events (see also [41]).

6.1 Simulation: single $I_0(E)$, single $\mu(E)$

A polyenergetic 120kVp source was modeled with a discrete incident energy spectrum, $I_0(E)$, spanning 0-150keV in 1keV steps. Projection sinograms were calculated in a planar, fan-beam geometry, with 1408 source angles per rotation and 768 detectors. The attenuation of the phantom is decomposed using single mass attenuation coefficient, $\mu(E)$, corresponding to water. The phantom shown in Figure 1 consists of several regions of different densities. The phantom body is made of Lucite and it holds two steel rods (one large and one small) and two aluminum rods. Images were reconstructed from the mean data (noiseless data) using filtered backprojection (FBP), an iterative deblurring algorithm (IDB) (see for example Snyder, et al. [35]), the monoenergetic version of the alternating minimization (AM) iterations described here, and the polyenergetic version of the AM algorithm from Section 4.1. The results are shown in Figure 2. All images presented here are based on attenuation values at 70 keV and are viewed in a [0.019 0.028] mm^{-1} window. The images are all relatively good for this noiseless case. In order to show the improvement of the new iterations quantitatively, a vertical profile through each image is shown in Figure 3 where x-axis represents row index and the y-axis is in units of mm^{-1} (location at which profiles are taken is indicated in panel (a)). The cupping artifact due to beam-hardening is evident in the first three images reconstructed using FBP, IDB algorithm, and the monoenergetic version of the AM algorithm, with a maximum overshoot of about 11%. The image reconstructed using the polyenergetic AM described in Section 4.1 does not exhibit the cupping artifact. Monotonic decrease of the objective function achieved with the polyenergetic AM algorithm is shown in Figure 4.

Reconstructions of the data including Poisson noise are shown in Figure 5. Vertical profiles through these reconstructions (taken at the same location as above) are shown in Figure 6. The total number of incident photons per source detector pair was set to 10^5, which may be typical of some low-dose scanning applications. The reduction of the cupping artifact seen in the noiseless case is evident here as well. As the number of iterations increases (necessary to eliminate streaks), the image estimates appear more grainy due to the lack of regularization.

Figure 7 presents reconstructions of data that include scatter, modeled as a constant, and Poisson noise. As before, the incident number of photons was 10^5, and the mean of the background events, $\beta(y)$, was set to 10. Vertical profiles in Figure 8 show the cupping artifact reduction achieved with the new algorithm. Also, we see in the bottom panel that accounting for scatter in our AM algorithm yields more accurate attenuation estimates (although again rather noisy here due to the relatively low total counts). In a simulation for which the number of incident photons was increased to 1.6×10^6 photons, the resulting images generally appear smoother as seen in Figure 9 which shows the reconstruction using the polyenergetic AM with scatter correction (displayed using the same window as in previous figures).

6.2 Simulation: multiple $I_0(E)$, multiple $\mu(E)$

In this simulation, we use the same phantom geometry as above but generate data using two constituent materials, styrene and calcium chloride solution, to approximate the attenuation of various phantom regions; the attenuation spectra are shown in the right panel of Figure 10. The component images are shown in Figure 11 (viewing windows are different for the two images so we include the colorbar for each). Here, two of the rods (one large and one small) simulate attenuation of the muscle tissue and the other two simulate Teflon, while the body of the phantom is kept as Lucite. We generate two noiseless data sets using different incident energy spectra. The first set is created using one incident energy spectrum $I_0(E)$ corresponding to 80kVp X-ray tube voltage and a second one using 140kVp tube voltage, as shown in the left panel of Figure 10. The AM algorithm iterations are easily modified to incorporate the two sets of data directly into the reconstruction (see [44]). Figure 12 shows the reconstruction of the component images obtained using the AM algorithm. The viewing windows for the two images are the same ones used in Figure 11.

There are many ways to evaluate the quality of the reconstructed images including: value of log-likelihood function or I-divergence achieved; quality of a combined image at a representative energy; errors in each component image; or a measure of mean performance relative to reconstruction noise level (for example, contrast-to-noise ratio) in each image or in a combined image. For the images shown, the reconstructed values are close to the true values used in the simulations. The highest attenuations in the first component (styrene) are achieved in the holes in the top left and the second to bottom on the right; the true values are 1.43. The reconstructed values have mean value 1.43 with standard deviation 0.01. In the background, we computed the average and standard deviation of both components over a rectangular region of size 213 by 41 just to the left of the holes, and not touching the edges. For the first component the average equals the true value of 1.14 with a standard deviation of 0.0003; for the second component the average is also the truth (0.0583) with standard deviation 0.0002. Qualitatively, there are some edge artifacts that are noticable in the second component. Also qualitatively, there is some streaking in the second component at a moderate level.

Stonestrom, *et al.* [36] use basis components corresponding to the photoelectric effect and Compton scattering. They and Williamson, *et al.* [42] describe methods to recover the relative partial densities using monoenergetic approximations to the scans and linear inversions.

7 Conclusions

Alternating minimization algorithms have been studied by many authors, following the work of I. Csiszár and G. Tusnády [7]. These algorithms are important in information theory [7, 28, 30, 31] and in image estimation [19, 29, 21, 4, 11, 38]. Our approach is based in part on the generalized iterative scaling algorithm derived rigorously by Darroch and Ratcliff [8] for minimizing relative entropy subject to linearity constraints. There are several general convergence properties of such algorithms that carry over, including monotonic decrease of the I-divergence (monotonic increase of the log-likelihood function); other properties do not carry over directly.

The convergence analysis for the algorithm here does not directly follow from results of Csiszár and Tusnády [7] or others (see for example, [5]) for several reasons. A primary reason is that $\mathbf{q}$ does not take values in a convex set. While the $c_i(x)$ do take values in a convex set and the I-divergence is convex in $c_i(x)$ and $\mathbf{p}$ separately, it is not always jointly convex in the pair $(\mathbf{p}, \{c_i(x)\})$ (of course it is jointly convex in $(\mathbf{p}, \mathbf{q})$). Additionally, the recommended algorithm performs only one update in the exponential family, thereby not reaching the global minimum over the exponential family prior to updating the member of the linear family.

It has been suggested that $h(y|x)\mu_i(E)$ could be renormalized to sum to one over (i, x) for every (y, E) (see Lemma 4 of Darroch and Ratcliff [8] for one basis of this suggestion), however this renormalization is not always possible and it may effectively separate the physical problem of interest (reconstruction of patient attenuation functions) from the mathematical problem. In particular, it does not suffice to obtain an optimal

$\mathbf{q}$–an estimate of $\{c_i(x)\}$ is required. That having been noted, the introduction of $Z_i(x)$ is motivated by this renormalization.

We have pursued algorithms for CT imaging in the presence of known high density attenuators that jointly estimate the pose of a known object and the patient attenuation [34, 41]; these algorithms may be adapted to the present framework.

Initial simulations have shown that the algorithm described here converges as predicted by the theory; images derived from noiseless monoenergetic data are comparable to those obtained using other algorithms. Additional simulations and more detailed performance analysis of this algorithm are given by Williamson, et al. [41].

Appendix A

In this appendix we prove Theorem 4 on monotonic convergence of the algorithm. The variational representation (9) implies that

$$I(\mathbf{d}\|\mathbf{g}(y:\hat{\mathbf{c}}^{(k)})) = I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}). \tag{45}$$

The update for $\hat{\mathbf{q}}$ at iteration $k+1$ satisfies

$$\begin{aligned} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}) - I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) &= \\ \sum_{y\in Y}\sum_{E}\Bigg[\hat{p}^{(k)}(y,E)\sum_i\sum_x h(y|x)\mu_i(E)[\hat{c}_i^{(k)}(x)-\hat{c}_i^{(k+1)}(x)] \\ +\hat{q}^{(k)}(y,E)\left(1-\exp\left(\sum_i\sum_x h(y|x)\mu_i(E)[\hat{c}_i^{(k)}(x)-\hat{c}_i^{(k+1)}(x)]\right)\right)\Bigg] \end{aligned} \tag{46}$$

In the iterations, the special case of $\hat{c}_i^{(k+1)}(x)=0$ that results from enforcing the nonnegativity constraint can be viewed as corresponding to selecting a higher value of $Z_i(x)$ which we denote by $\tilde{Z}_i^{(k)}(x)$. If it is also true that $\hat{c}_i^{(k)}(x)=0$, then this higher value may be taken to be $\tilde{Z}_i^{(k)}(x)=\infty$; define the set of indices that exclude those such that $\hat{c}_i^{(k)}(x)=\hat{c}_i^{(k+1)}(x)=0$ as $\mathcal{I}^{(k)}$. Using $\tilde{Z}_i^{(k)}(x)$, we have the equality

$$\hat{c}_i^{(k)}(x)-\hat{c}_i^{(k+1)}(x) = \frac{1}{\tilde{Z}_i^{(k)}(x)}\ln\frac{\tilde{b}_i^{(k)}(x)}{\hat{b}_i^{(k)}(x)}. \tag{47}$$

Now note that

$$\begin{aligned} \sum_{y\in Y}\sum_E \hat{q}^{(k)}(y,E) &\geq \sum_{y\in Y}\sum_E \hat{q}^{(k)}(y,E)\sum_x\sum_i h(y|x)\mu_i(E)\frac{1}{Z_i(x)} \\ &= \sum_x\sum_i \frac{\hat{b}_i^{(k)}(x)}{Z_i(x)} \\ &\geq \sum_x\sum_i \frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)}. \end{aligned} \tag{48}$$

Because $\tilde{Z}_i^{(k)}(x)\geq Z_i(x)$, the convex decomposition lemma applied to (46) holds using $\tilde{Z}_i^{(k)}(x)$ so

$$\begin{aligned} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}) - I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) &\geq \\ \sum_{(i,x)\in\mathcal{I}^{(k)}}\left[\tilde{b}_i^{(k)}(x)[\hat{c}_i^{(k)}(x)-\hat{c}_i^{(k+1)}(x)] - \hat{b}_i^{(k)}(x)\frac{1}{\tilde{Z}_i^{(k)}(x)}\exp\left(\tilde{Z}_i^{(k)}(x)[\hat{c}_i^{(k)}(x)-\hat{c}_i^{(k+1)}(x)]\right)\right] \\ +\sum_{y\in Y}\sum_E \hat{q}^{(k)}(y,E) \quad . \end{aligned} \tag{49}$$

Substituting (47) and (48) into (49) yields

$$I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}) - I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) \quad \geq$$
$$\sum_{(i,x)\in\mathcal{I}^{(k)}} \left[\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \ln \frac{\tilde{b}_i^{(k)}(x)}{\hat{b}_i^{(k)}(x)} - \frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} + \frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \right] \quad . \tag{50}$$

More succinctly,

$$\begin{aligned} I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k)}) - I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) &\geq I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \| \frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \right) \\ &\geq 0, \end{aligned} \tag{51}$$

where we set $0 \ln \frac{0}{0} = 0$.

The triangle equality in the first variable for linear families yields

$$I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{q}}^{(k+1)}) \quad = \quad I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{p}}^{(k+1)}) + I(\hat{\mathbf{p}}^{(k+1)}\|\hat{\mathbf{q}}^{(k+1)}). \tag{52}$$

Putting together (45), (51), and (52), we have

$$I(\mathbf{d}\|\mathbf{g}(y:\hat{\mathbf{c}}^{(k)})) - I(\mathbf{d}\|\mathbf{g}(y:\hat{\mathbf{c}}^{(k+1)})) \quad \geq \quad I(\hat{\mathbf{p}}^{(k)}\|\hat{\mathbf{p}}^{(k+1)}) + I\left(\frac{\tilde{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \| \frac{\hat{b}_i^{(k)}(x)}{\tilde{Z}_i^{(k)}(x)} \right). \tag{53}$$

That is, the iterations monotonically decrease the log-likelihood function.

The right side of (53) equals zero if and only if $\tilde{b}_i^{(k)}(x)/\tilde{Z}_i^{(k)}(x) = \hat{b}_i^{(k)}(x)/\tilde{Z}_i^{(k)}(x)$ for all $(i,x) \in \mathcal{I}^{(k)}$. For all (x,i) such that $\hat{c}_i^{(k)}(x) > 0$, this condition implies $\tilde{b}_i^{(k)}(x) = \hat{b}_i^{(k)}(x)$. For all (x,i) such that $\hat{c}_i^{(k)}(x) = \hat{c}_i^{(k+1)}(x) = 0$, we have from the iterations that $\tilde{b}_i^{(k)}(x) \geq \hat{b}_i^{(k)}(x)$. The case $\hat{c}_i^{(k+1)}(x) > \hat{c}_i^{(k)}(x) = 0$ is not possible if there is equality. Thus, the Kuhn- Tucker conditions are satisfied. Under these conditions, $\hat{c}_i^{(k)}(x) = \hat{c}_i^{(k+1)}(x)$ for all (i,x) and we have equality in (33).

This proves the inequality (33) and the theorem statement.

Acknowledgments. We are grateful to the reviewers and the associate editor who provided very detailed comments about the paper. R. J. Murphy, D. G. Politte, D. L. Snyder, B. R. Whiting, and J. F. Williamson contributed to ideas and simulations in support of this work; they, J. Fessler, and N. Singla provided many helpful comments on the manuscript.

References

[1] S. Ahn, J. A. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," *IEEE Trans. Med. Imaging*, 22(5):613-26, May 2003.

[2] J. Benac, "Alternating Minimization Algorithms for X-Ray Computed Tomography: Multigrid Acceleration and Dual Energy Application," D.Sc. Thesis, Department of Electrical and Systems Engineering, Washington University, St. Louis, MO, May 2005.

[3] J. Browne and A. R. De Pierro, "A Row-Action Alternative to the EM Algorithm for Maximizing Likelihoods in Emission Tomography," *IEEE Trans. on Med. Imaging*, 15(5):687-699, 1996.

[4] C. L. Byrne, "Iterative Image Reconstruction Algorithms Based on Cross-Entropy Minimization," *IEEE Transactions on Image Processing,* vol. 2, no. 1, pp. 96-103, 1993.

[5] C. Byrne and Y. Censor, "Proximity Function Minimization Using Multiple Bregman Projections, with Applications to Split Feasibility and Kullback-Leibler Distance Minimization," *Annals of Operations Research,* vol. 105, pp. 77-98, 2001.

[6] I. Csiszár, "Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inference for Linear Inverse Problems," *The Annals of Statistics*, vol. 19, pp. 2032-2066, 1991.

[7] I. Csiszár and G. Tusnády, "Information Geometry and Alternating Minimization Procedures," *Statistical Decisions*, Suppl. issue no. 1, pp. 205-207, 1984.

[8] J. N. Darroch and D. Ratcliff, "Generalized Iterative Scaling for Log-Linear Models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470-1480, 1972.

[9] B. De Man, J. Nuyts, P. Dupont, G. Marchal, and P. Suetens, "An Iterative Maximum-Likelihood Polychromatic Algorithm for CT," *IEEE Transactions on Medical Imaging,* vol. 20, no. 10, pp. 999-1008, Oct. 2001.

[10] A. R. De Pierro, "On the Relation Between the ISRA and EM Algorithm for Positron Emission Tomography," *IEEE Transactions on Medical Imaging,* vol. 12, pp. 328-333, 1993.

[11] A. R. De Pierro, "A Modified Expectation Maximization Algorithm for Penalized Likelihood Estimation in Emission Tomography," *IEEE Transactions on Medical Imaging,* vol. 14, pp. 132-137, 1995.

[12] I. A. Elbakri and J. A. Fessler, "Statistical X-ray Computed Tomography Image Reconstruction with Beam Hardening Correction," *Proc. SPIE 4322, Medical Imaging 2001: Image Proc.*, 2001.

[13] I. A. Elbakri and J. A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," *IEEE Transactions on Medical Imaging,* vol. 21, no. 2, pp. 89-99, Feb. 2002.

[14] H. Erdogan and J. A. Fessler, "Monotonic Algorithms for Transmission Tomography," *IEEE Trans. on Med. Imaging*, 18(9):801-814, 1999.

[15] H. Erdogan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, 44(11):2835-2851, 1999.

[16] J. A. Fessler, "Statistical image reconstruction methods for transmission tomography," in *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis*, M. Sonka and J. Michael Fitzpatrick, eds. SPIE, Bellingham, pp. 1-70, 2000.

[17] H. M. Hudson and R. S. Larkin, "Accelerated Image Reconstruction Using Ordered Subsets of Projection Data," *IEEE Transactions on Medical Imaging,* vol. 13, no. 4, pp. 601-609, 1994.

[18] C. Kamphuis and F. J. Beekman, "Accelerated Iterative Transmission CT Reconstruction Using an Ordered Subsets Convex Algorithm," *IEEE Trans. on Med. Imaging*, 17(6):1101-1105, 1998.

[19] K. Lange, "Convergence of EM Image Reconstruction Algorithms with Gibbs Smoothing," *IEEE Trans. on Med. Imaging*, 9(4):439-446, 1990.

[20] K. Lange and R. Carson, "EM Reconstruction Algorithms for Emission and Transmission Tomography," *Journal of Computer Assisted Tomography*, vol. 8, no. 2, pp. 306-316, 1984.

[21] K. Lange and J. A. Fessler, "Globally Convergent Algorithms for Maximum a Posteriori Transmission Tomography," *IEEE Transactions on Image Processing,* vol. 4, no. 10, pp. 1430-1438, 1995.

[22] M. I. Miller and D. L. Snyder, "The Role of Likelihood and Entropy in Incomplete-Data Problems: Applications to Estimating Point-Process Intensities and Toeplitz and Constrained Covariances," *Proceedings of the IEEE*, vol. 75, pp. 892-907, 1987.

[23] J. Nuyts, B. De Man, P. Dupont, M. Defrise, P. Suetens, and L. Mortelmans, "Iterative Reconstruction for Helical CT: A Simulation Study," *Phys. Med. Biol.*, vol. 43, pp. 729-737, 1998.

[24] R. J. Murphy, "Transmission Tomographic Image Reconstsruction Using Alternating Minimization Algorithms," D.Sc. Thesis, Department of Electrical and Systems Engineering, Washington University, St. Louis, MO, May 2004.

[25] R. J. Murphy, S. Yan, J. A. OSullivan, D. L. Snyder, B. R. Whiting, D. G. Politte, G. Lasio, and J. F. Williamson, "Pose Estimation of Known Objects During Transmission Tomographic Image Reconstruction," submitted to the *IEEE Trans. Medical Imaging*, 2006.

[26] J. M. Ollinger, "Maximum-Likelihood Reconstruction of Transmission Images in Emission Computed Tomography via the EM Algorithm," *IEEE Trans. on Med. Imaging*, 13(1):89-101, 1994.

[27] J. A. O'Sullivan, "Roughness Penalties on Finite Domains," *IEEE Transactions on Information Theory*, vol. 4, no. 9, pp. 1258-1268, 1995.

[28] J. A. O'Sullivan, "Alternating Minimization Algorithms: from Blahut-Arimoto to Expectation-Maximization," in A. Vardy, Ed., *Codes, Curves, and Signals: Common Threads in Communications,* 1998, pp. 173-192.

[29] J. A. O'Sullivan, R. E. Blahut, and D. L. Snyder, "Information Theoretic Image Formation," invited paper for the special issue of the *IEEE Transactions on Information Theory* in honor of the 50th anniversary of C. E. Shannon's 1948 paper, vol. 44, no. 6, pp. 2094-2123, October 1998. Also in *Information Theory: 50 Years of Discovery*, S. Verdu and S. W. McLaughlin, Eds., pp. 50-79, 2000.

[30] J. A. O'Sullivan, "Information Geometry and the Information Value Decomposition," *Proceedings IEEE 2000 International Symposium on Information Theory*, Sorrento, Italy, p. 491, June 2000.

[31] J. A. O'Sullivan, "Iterative Algorithms for Maximum-Likelihood Sequence Detection," in R. E. Blahut, and R. Koetter, eds., *Codes, Graphs, and Systems*, Kluwer Academic, Boston, pp. 137-156, 2002.

[32] J. Benac, J. A. O'Sullivan, and J. F. Williamson, "Alternating Minimization Algorithm for Dual Energy X-Ray CT," *Proc. IEEE International Symp. Biomedical Imaging,* pp. 579-582, Arlington, VA, April 2004.

[33] D. L. Snyder, J. A. OSullivan, R. J. Murphy, D. G. Politte, B. R. Whiting, and J. F. Williamson, "Maximum likelihood image reconstruction for transmission tomography when projection data are incomplete," in preparation, 2006.

[34] D. L. Snyder, J. A. O'Sullivan, B. R. Whiting, R. J. Murphy, J. Benac, J. A. Cataldo, D. G. Politte, and J. F. Williamson, "Deblurring subject to Nonnegativity Constraints When Known Functions are Present, with Application to Object-Constrained Computerized Tomography," *IEEE Transactions on Medical Imaging,* Vol. 20, No. 10, PP. 1009-1017, October 2001.

[35] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan, "Deblurring Subject to Nonnegativity Constraints," *IEEE Transactions on Signal Processing,* Vol. 40, No. 5, May 1992, pp. 1143-1150.

[36] J. P. Stonestrom, R. E. Alvarez, and A. Macovski, "A Framework for Spectral Artifact Corrections in X-Ray CT," *IEEE Transactions on Biomedical Engineering*, vol. BME-28, no. 2, pp. 128-141, Feb. 1981.

[37] P. Sukovic and N. H. Clinthorne, "Penalized Weighted Least-Square Image Reconstruction for Dual Energy X-Ray Transmission Tomography," *IEEE Transactions on Medical Imaging*, vol. 19, no. 11, pp. 1075-1081, November 2000.

[38] G. Wang, D. L. Snyder, J. A. O'Sullivan, and M. W. Vannier, "Iterative Deblurring for CT Metal Artifact Reduction," *IEEE Transactions on Medical Imaging*, vol. 15, pp. 657-664, October 1996.

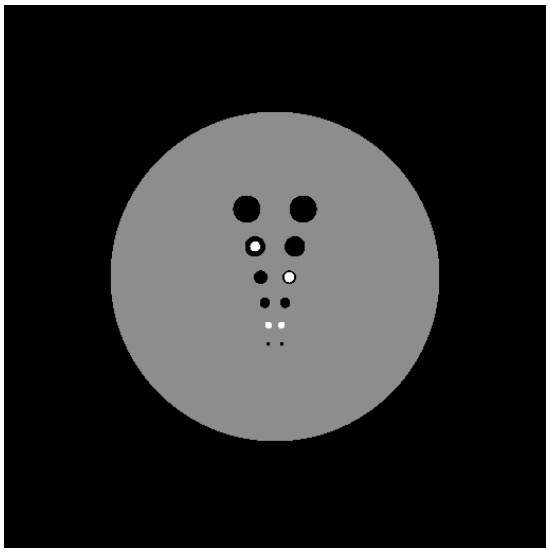


Figure 1: The phantom truth image.

[39] B. R. Whiting, "Signal Statistics of X-ray Computed Tomography," *Proc. SPIE Conference on Medical Imaging 2002: Physics of Medical Imaging*, vol. 4682 (L. Antonuk and M. Yaffe, Eds.), San Diego, CA, Feb. 2002.

[40] B. R. Whiting, P. Massoumzadeh, O. A. Earl, J. A. O'Sullivan,D. L. Snyder, and J. F. Williamson, "X-ray Computed Tomography Signal Properties," submitted to *Medical Physics*, January 2006.

[41] J. F. Williamson, B. R. Whiting, J. Benac, R. J. Murphy, G. J. Blaine, J. A. O'Sullivan, D. G. Politte, and D. L. Snyder, "Prospects for Quantitative Computed Tomography Imaging in the Presence of Foreign Metal Bodies Using Statistical Image Reconstruction," *Med. Phys.* Vol. 29, No. 10, pp. 2404-2418, October 2002.

[42] J. F. Williamson, S. Li, S. Devic, B. R. Whiting, and F. A. Lerma, "On Two-Parameter Representations of Photon Cross Section: Application to Dual Energy CT Imaging," Submitted for publication to *Med. Phys.*, November 2003.

[43] C. H. Yan, R. T. Whalen, G. S. Beaupre, S. Y. Yen, and S. Napel, "Reconstruction Algorithm for Polychromatic CT Imaging: Application to Beam Hardening Correction," *IEEE Trans. on Med. Imaging*, vol. 19, no. 1, pp. 1-11, 2000.

[44] J. A. O'Sullivan, , J. Benac, and J. F. Williamson, "Alternating minimization algorithm for dual energy X-ray CT," *Proc. IEEE International Symposium on Biomedical Imaging*, 2004.

[45] H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part 3*, Krieger Publishing Company, 1992.

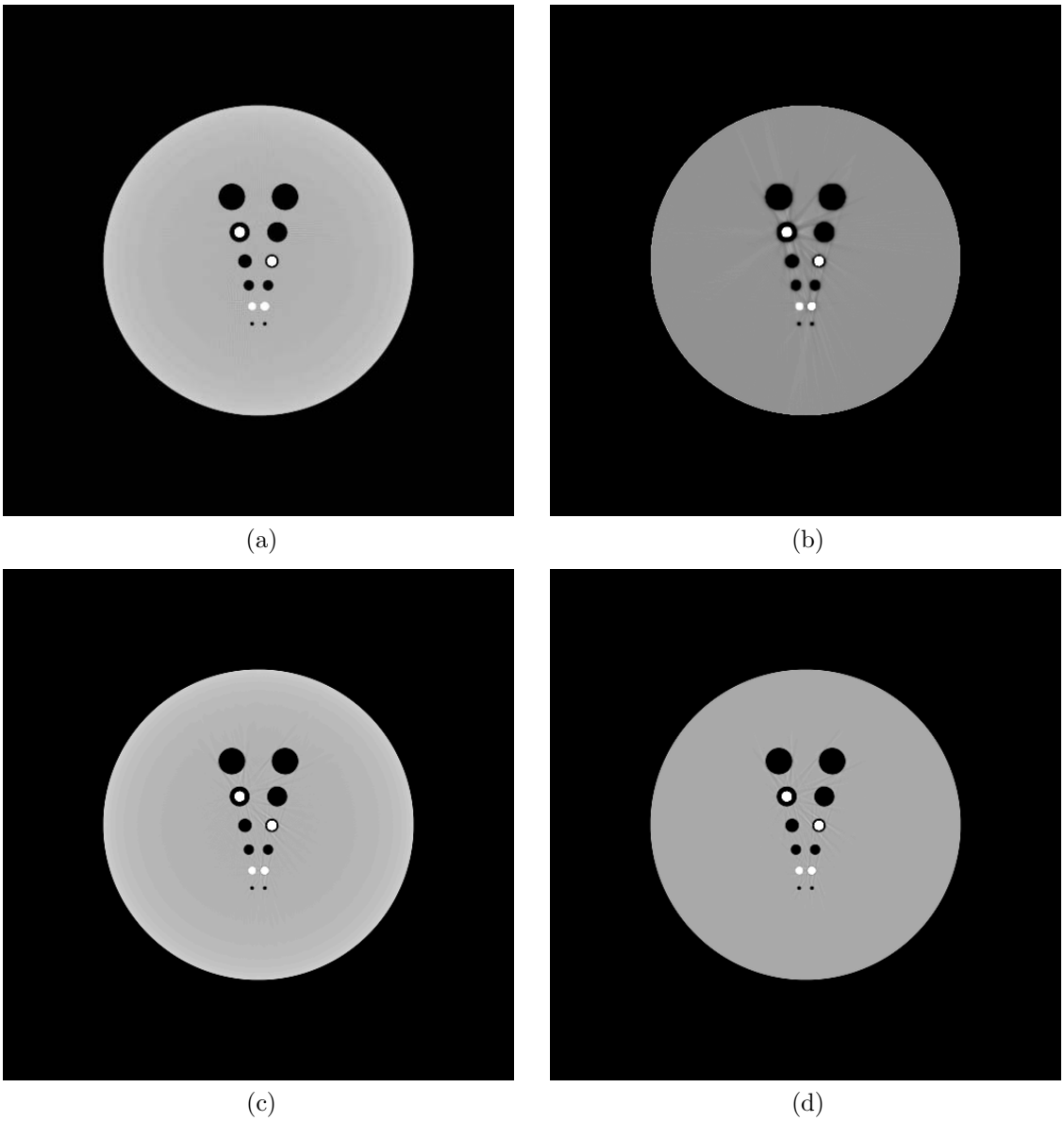
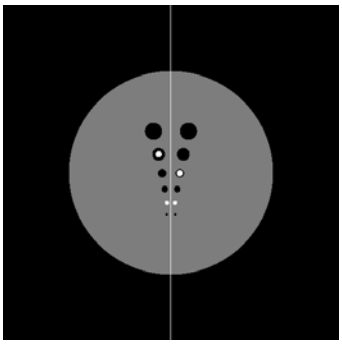
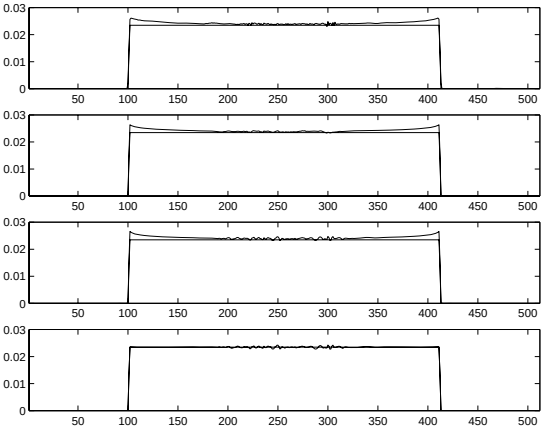


Figure 2: Noiseless data reconstructions comparison: (a) filtered backprojection (FBP), (b) iterative deblurring algorithm (IDB), (c) monoenergetic alternating minimization (AM) algorithm, and (d) polyenergetic AM algorithm.



(a)



(b)

Figure 3: (a) Truth image with marked column where profiles will be taken, (b) Profiles through the reconstructions versus the truth (top to bottom): FBP, IDB, monoenergetic AM algorithm, and polyenergetic AM algorithm.

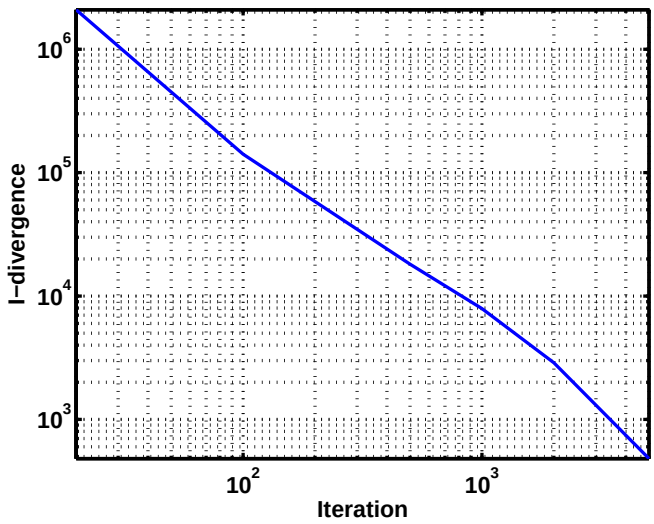


Figure 4: Plot of I-divergence vs iteration number for polyenergetic AM (noiseless data).

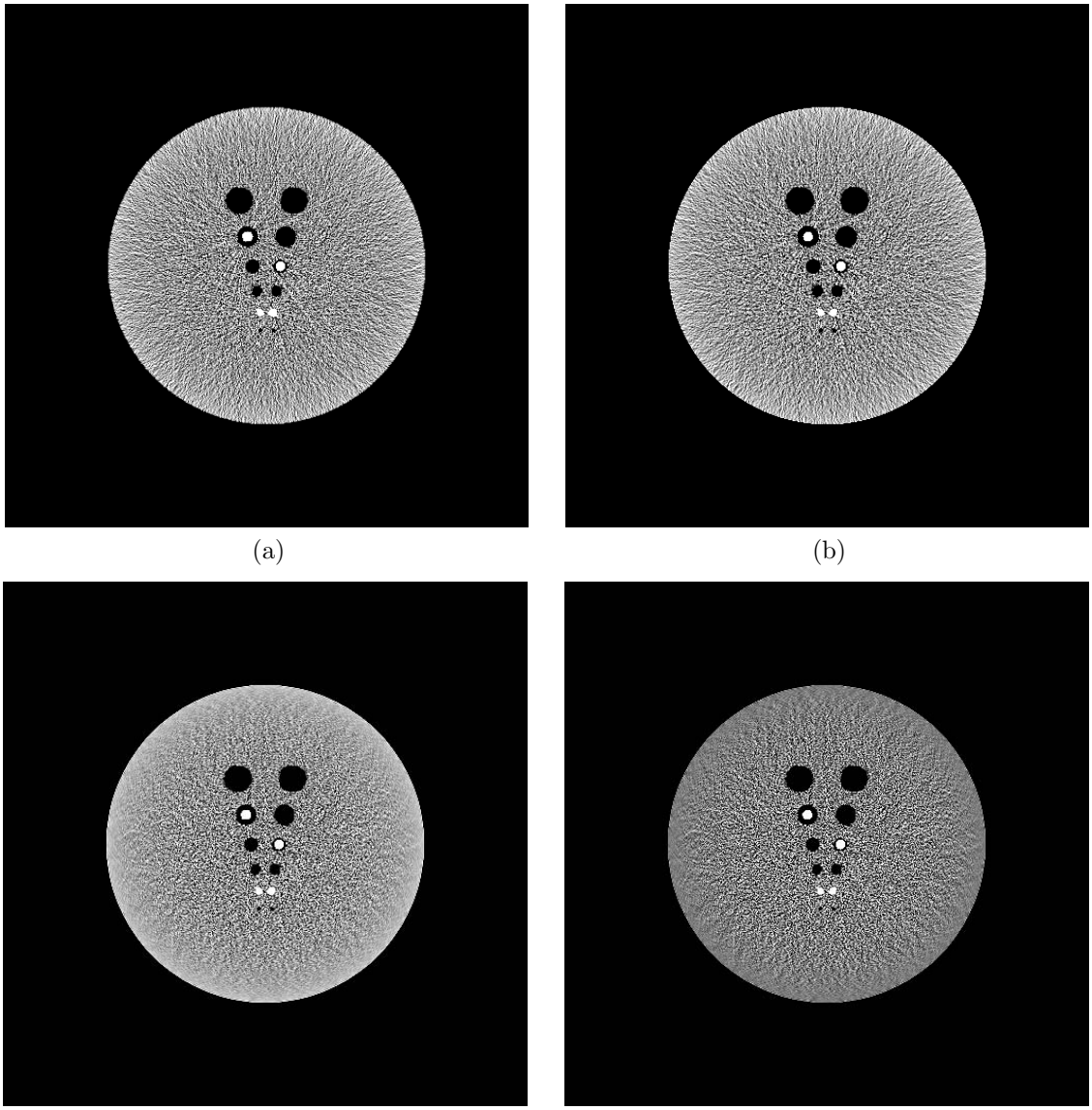


Figure 5: Noisy data reconstructions comparison: (a) filtered backprojection (FBP), (b) iterative deblurring algorithm (IDB), (c) monoenergetic AM algorithm, and (d) polyenergetic AM algorithm.

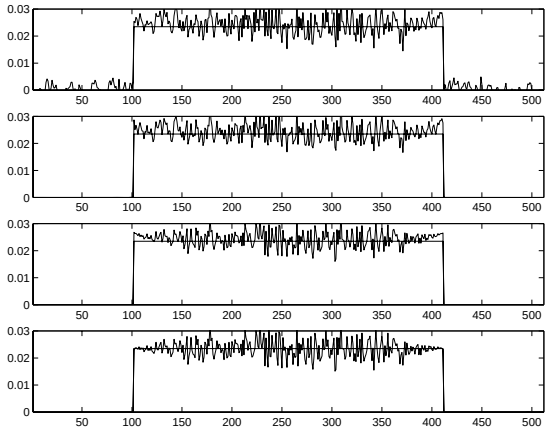


Figure 6: Vertical cut through the reconstructions versus the truth (top to bottom): FBP, IDB, monoenergetic AM algorithm, and polyenergetic AM algorithm.

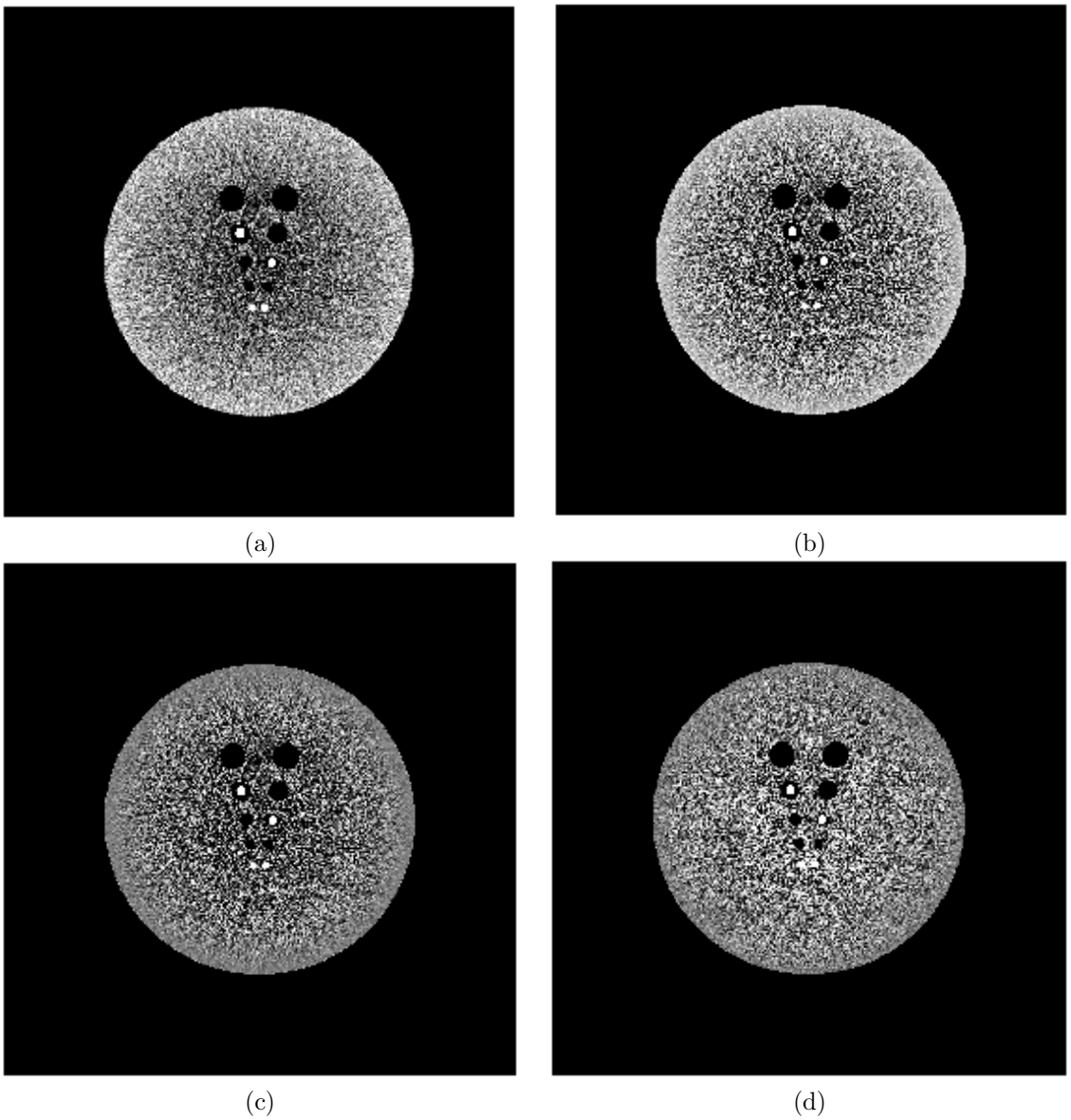


Figure 7: Noisy data reconstructions comparison: (a) FBP, (b) monoenergetic AM algorithm, (c) polyenergetic AM without scatter correction, and (d) polyenergetic AM with scatter correction.

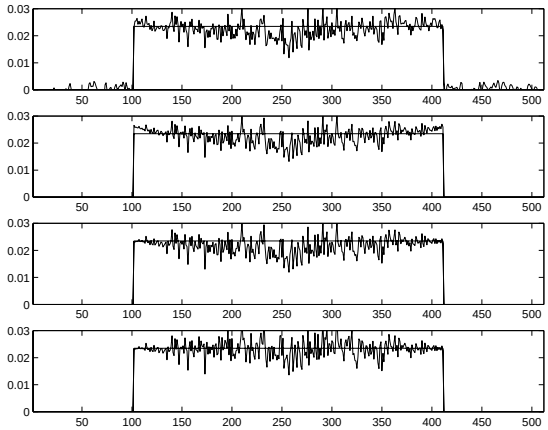


Figure 8: Vertical profiles through the reconstructions versus the truth (from top to bottom): FBP, IDB, monoenergetic AM algorithm, polyenergetic AM algorithm without scatter correction, and polyenergetic AM with scatter correction.

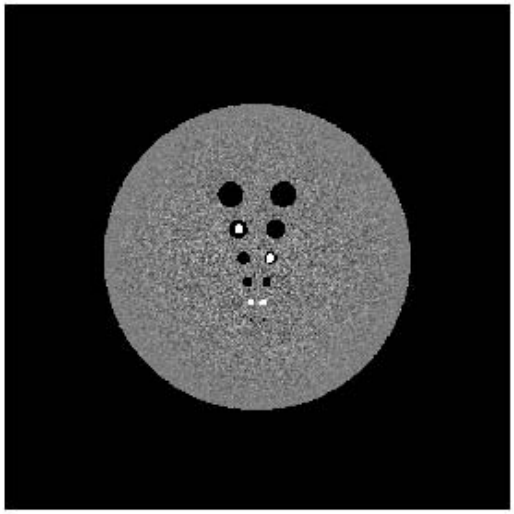


Figure 9: Polyenergetic AM with scatter correction on data created using higher value of $I_0(y)$.

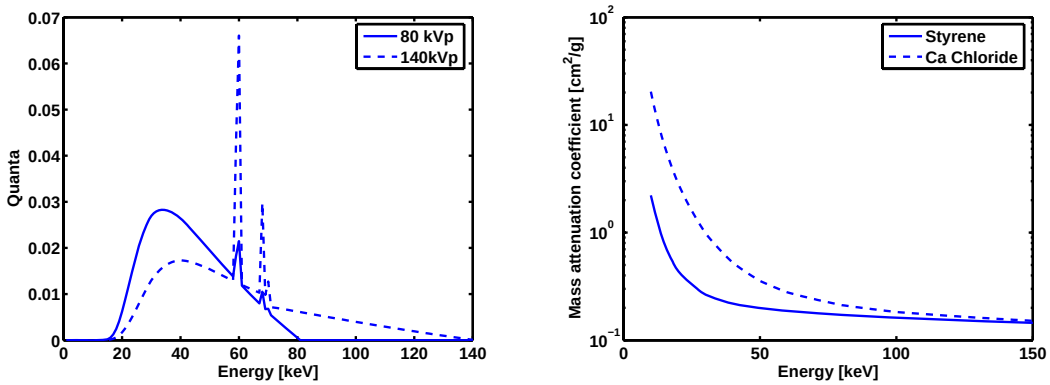


Figure 10: The left panel shows the incident energy spectra for 80kVp and 140kVp. The right panel shows the consituent spectra for styrene and CaCl used in the simulations.

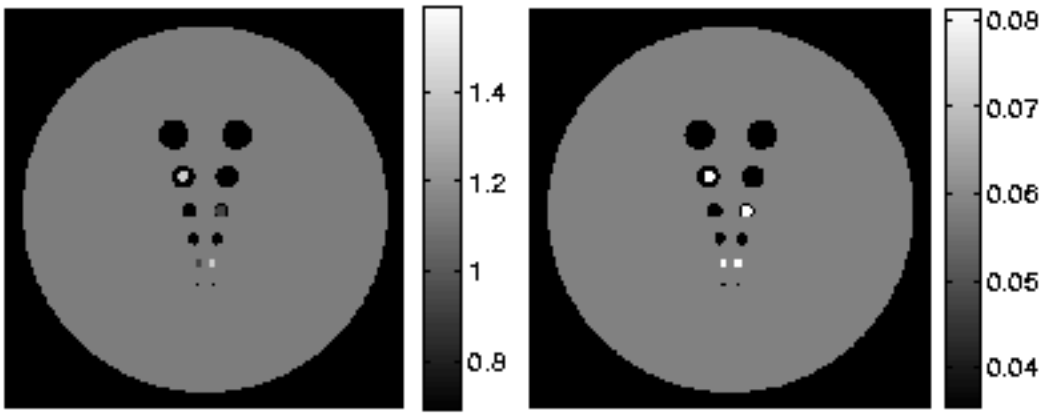


Figure 11: Component images used to simulated dual-energy data.

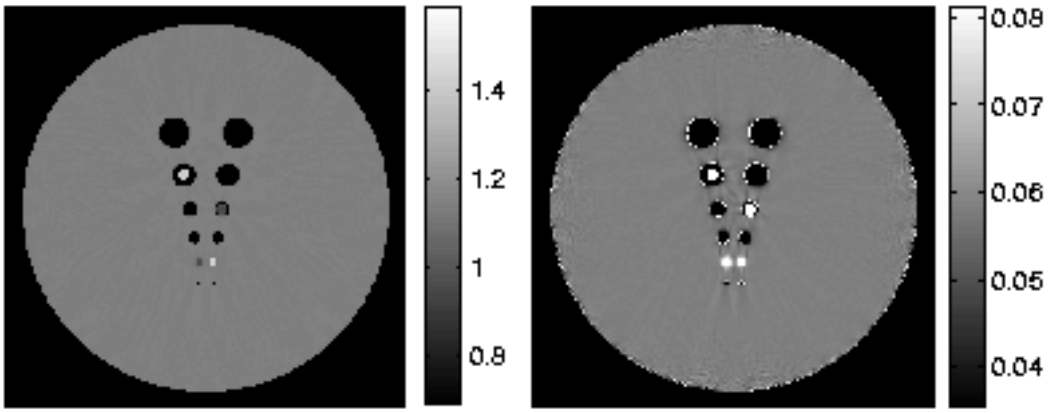


Figure 12: Dual-energy AM reconstruction of the component images after 1700 iterations.