Fast Quasi-Newton Algorithms for Penalized Reconstruction in Emission Tomography and Further Improvements via Preconditioning

Yu-Jung Tsai, *Student Member, IEEE*, Alexandre Bousse, Matthias J. Ehrhardt,
Charles W. Stearns, *Fellow, IEEE*, Sangtae Ahn, *Member, IEEE*, Brian F. Hutton, *Senior Member, IEEE*,
Simon Arridge, and Kris Thielemans, *Senior Member, IEEE*

***Abstract*— This paper reports on the feasibility of using a quasi-Newton optimization algorithm, limited-memory Broyden-Fletcher-Goldfarb-Shanno with boundary constraints (L-BFGS-B), for penalized image reconstruction problems in emission tomography (ET). For further acceleration, an additional preconditioning technique based on a diagonal approximation of the Hessian was introduced. The convergence rate of L-BFGS-B and the proposed preconditioned algorithm (L-BFGS-B-PC) was evaluated with simulated data with various factors, such as the noise level, penalty type, penalty strength and background level. Data of three ^{18}F-FDG patient acquisitions were also reconstructed. Results showed that the proposed L-BFGS-B-PC outperforms L-BFGS-B in convergence rate for all simulated conditions and the patient data. Based on these results, L-BFGS-B-PC shows promise for clinical application.**

***Index Terms*— Emission tomography, penalized reconstruction, L-BFGS-B, preconditioning.**

I. Introduction

EMISSION tomography (ET) allows non-invasive observation of metabolic processes *in vivo*. With adequate image processing and analysis methods, it is valuable for the diagnosis of many diseases. In current clinical practice, most applications of ET are based on visual interpretation. With the expansion of its potential clinical application, such as disease follow-up and therapy monitoring [1]–[3], there is increased interest in precise quantification of the images. The reconstructed images are therefore expected to accurately represent the tracer concentration.

Due to their ability to include modeling of the imaging physics and statistics, iterative reconstruction algorithms have become the method of choice for pursuing both good visual quality and high quantitative accuracy, most often based on maximum-likelihood (ML) estimation. However, image reconstruction using ML estimation is an ill-conditioned problem, resulting in noise amplification as iterations increase [4]. In practice, the noise can be controlled by early termination of the iterative process, at the expense of quantitative accuracy [5], or by incorporation of a penalty term [6], [7]. One of the most widely used methods for incorporating a penalty term is the one-step-late (OSL) approach [8]. Although it can be applied with any differentiable penalty function, the algorithm can be unstable and divergent for large penalty strength [9]. Modified ML-EM algorithms [10] or separable paraboloidal surrogates (SPS) [11] can directly incorporate the penalty term into a closed-form update of the image without suffering from convergence issues. However, the application of both strategies is limited by the need to find a convex surrogate function.

Another alternative is to employ the generic steepest-descent optimization algorithm to find the local solution along the gradient of the penalized likelihood function by using a line search. With a good line search algorithm, steepest descent can show fast initial convergence rate but often slows down while approaching the final solution as the direction defined by the gradient can lead to a zigzag path to the solution for ill-conditioned problems. Instead of using merely the gradient, Newton's method [12] defines a better search direction with the help of the Hessian matrix. However, the Hessian in large scale problems is usually too large to calculate or store in memory and may be non-invertible. To overcome this, quasi-Newton algorithms that use approximations for the Hessian were therefore developed.

Manuscript received September 25, 2017; revised December 11, 2017; accepted December 18, 2017. Date of publication December 25, 2017; date of current version April 2, 2018. This work was supported in part by GE Healthcare and in part by the NIHR-funded UCLH Biomedical Research Centre. The work of M. J. Ehrhardt was supported in part by the Leverhulme Trust Project—Breaking the non-convexity barrier, EPSRC, under Grant EP/M00483X/1 and Grant EP/N014588/1, in part by the Cantab Capital Institute for the Mathematics of Information, and in part by CHiPS (Horizon 2020 RISE Project Grant). *(Corresponding author: Yu-Jung Tsai.)*

Y.-J. Tsai, A. Bousse, and K. Thielemans are with the Institute of Nuclear Medicine, University College London, London NW1 2BU, U.K. (e-mail: yu-jung.tsai.14@ucl.ac.uk).

M. J. Ehrhardt is with the Department for Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, U.K.

C. W. Stearns is with MICT Engineering, GE Healthcare, Waukesha, WI 53188 USA.

S. Ahn is with GE Global Research, Niskayuna, NY 12309 USA.

B. F. Hutton is with the Institute of Nuclear Medicine, University College London, London NW1 2BU, U.K., and also with the Centre for Medical Radiation Physics, University of Wollongong, Wollongong, NSW 2522, Australia.

S. Arridge is with the Department of Computer Science, University College London, London WC1E 6BT, U.K.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2017.2786865

A popular example is the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [13], [14], which approximates the inverse of the Hessian based on the gradient information in the last few iterations. L-BFGS has been extended to allow box constraints on the variables that are to be estimated (L-BFGS-B) [15], [16]. Since the amount of memory the algorithm requires can be controlled by the user and scales linearly with the dimension of the problem, the algorithm has become the most popular quasi-Newton method for optimizing nonlinear problems [17]. It is widely used in machine learning but not yet in penalized-ML (PML) image reconstruction. As L-BFGS-B constructs approximates of the inverse Hessian by using only the gradient information, the algorithm should be able to handle any differentiable penalty term. This enables the incorporation of many non-convex penalty functions, such as the recently developed parallel level sets (PLS) [18] and the joint entropy priors [19]. Its wide applicability together with fast convergence rate make L-BFGS-B a promising candidate for a general-purpose optimization algorithm for PML image reconstruction.

In an initial study [20], we observed that L-BFGS-B can converge several times faster than OSL-EM [8] and relaxed SPS [21]. However, some issues were found that made the use of L-BFGS-B difficult for image reconstruction in ET as the observed convergence rate was dependent on image and data scale. This paper concentrates on improving the performance of L-BFGS-B by introducing better initialization and additional diagonal preconditioning. Previously, Kaplan et al. used L-BFGS-B with a preconditioner for accelerating simultaneous estimation of activity and attenuation distributions in single-photon emission computed tomography (SPECT) [22]. A constant value was chosen as the preconditioner to rescale the activity estimate. The algorithm was able to show a faster convergence rate in most cases when both the transformed activity and attenuation were in a similar scale. However, since the scale of the activity varies with application and individual dataset, the preconditioner had to be tuned accordingly by trial and error. Here, we use a more general diagonal preconditioner based on the second partial derivative of the objective function. With the help of the extra information, the penalized reconstruction problem is transformed to a better-conditioned form which is then incorporated into the L-BFGS-B optimization process. We denote the resulting algorithm as L-BFGS-B-PC.

A brief description of the PML optimization problem and the penalty terms used is given in section II. Section III provides an insight on the L-BFGS-B approach as well as the derivation of L-BFGS-B-PC. The evaluation methods used in this study are described in section IV. In section V, evaluations of L-BFGS-B and L-BFGS-B-PC are performed using digital simulations. The feasibility of applying both algorithms in a clinical context is then assessed on three patient data sets. Discussion and conclusions are presented in sections VI and VII, respectively. This paper expands on initial results previously presented by our group [23].

II. Penalized Maximum-Likelihood Image Reconstruction

A. Objective Function

In ET, the measured data $\boldsymbol{g} \in \mathbb{R}^I$ given a tracer distribution $\boldsymbol{f} \in \mathbb{R}^J$ can be described using a Poisson model:

$$\boldsymbol{g} \sim \text{Poisson}(\bar{\boldsymbol{g}}(\boldsymbol{f})), \quad \bar{\boldsymbol{g}} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{A}$ is the $I \times J$ system matrix and $\boldsymbol{n} \in \mathbb{R}^I$ is the expected background events vector, such as scatter and random coincidences. Each element of $\boldsymbol{A}$, A_{ij}, denotes the probability that an emission from voxel j is detected by bin i. Taking the logarithm and omitting terms independent of $\boldsymbol{f}$, the log-likelihood function of $\boldsymbol{g}$ is:

$$L(\boldsymbol{g}|\boldsymbol{f}) = \sum_i g_i \log \bar{g}_i(\boldsymbol{f}) - \bar{g}_i(\boldsymbol{f}). \tag{2}$$

Maximizing L is equivalent to minimizing $-L$. The optimization therefore becomes a minimization problem. In the rest of the paper, we refer to the optimization as a minimization problem. Instead of optimizing (2), PML image reconstruction minimizes the objective function Φ, which consists of the negative likelihood $-L$ and the penalty function R with a parameter β controlling its strength:

$$\Phi(\boldsymbol{f}) = -L(\boldsymbol{g}|\boldsymbol{f}) + \beta R(\boldsymbol{f}). \tag{3}$$

The optimization of the problem can then be addressed as:

$$\hat{\boldsymbol{f}} = \arg\min_{\boldsymbol{f} \geq \boldsymbol{0}} \Phi(\boldsymbol{f}). \tag{4}$$

Note that a positivity constraint is enforced on $\boldsymbol{f}$, as it represents radioactivity concentration.

B. Penalty Functions

Several penalty functions can be used to control noise propagation [24]–[26]. In this study, for simplicity, we use Gibbs-type penalties, which penalize the difference between voxels in a given neighborhood $\mathcal{N}$:

$$R(\boldsymbol{f}) = \frac{1}{2}\sum_k \sum_{j \in \mathcal{N}_k} \omega_{jk}\varphi(f_j - f_k) \tag{5}$$

where ω_{jk} indicates the weight between voxel j and its neighboring voxel k. We used two potential functions φ: the quadratic penalty (QP) and the rescaled log-cosh penalty (LP):

$$\varphi_{\text{QP}}(x) = x^2, \quad \varphi_{\text{LP}}(x) = \frac{1}{\rho^2}\log(\cosh(\rho x)) \tag{6}$$

where ρ is a scalar controlling the edge-preservation property of φ_{LP}. The factor $1/\rho^2$ is derived from the second derivative of φ_{LP} for normalization such that both priors behave similarly for small $|x|$. Note that for penalties as defined in (5) and (6), Φ is strictly convex [21].

III. Algorithms

A. L-BFGS-B

In this section, we describe the main ideas behind L-BFGS-B. More detail can be found in [15] and [27].

1) Unconstrained Optimization: Given the objective function Φ and current estimate $\boldsymbol{f}_t$ at iteration t, a polynomial approximation of Φ in the neighborhood of $\boldsymbol{f}_t$ is

$$q_t(\boldsymbol{f}) = \Phi(\boldsymbol{f}_t) + \boldsymbol{v}_t^\top \nabla\Phi(\boldsymbol{f}_t) + \frac{1}{2}\boldsymbol{v}_t^\top \boldsymbol{B}_t^{-1}\boldsymbol{v}_t. \tag{7}$$

where $\boldsymbol{v}_t = \boldsymbol{f} - \boldsymbol{f}_t$ and $\boldsymbol{B}_t$ is an approximation of the inverse of the Hessian matrix $\boldsymbol{H}$ at $\boldsymbol{f}_t$. The latter can be computed using L-BFGS using limited memory. The algorithm does not store $\boldsymbol{B}_t$ directly, but represents it by a pair of lower-dimensional correction matrices, which record the change of the update and the gradient of the objective function in the last few iterations, in order to compute the matrix/vector products with $\boldsymbol{B}_t$ efficiently [28]. A description of the construction of $\boldsymbol{B}_t$ is given in Appendix.

When $\boldsymbol{B}_t$ is positive definite, and ignoring the positivity constraint, q_t has a unique minimizer $\boldsymbol{f}^\star$:

$$\boldsymbol{f}^\star = \boldsymbol{f}_t - \boldsymbol{B}_t \nabla\Phi(\boldsymbol{f}_t). \tag{8}$$

Since the polynomial approximation (7) is local, $\boldsymbol{f}^\star$ cannot be used as an update for the minimization of Φ. Instead, we seek an update $\boldsymbol{f}_{t+1}$ along the line segment $\{\boldsymbol{f}_t + \alpha\boldsymbol{v}_t^\star,\ \alpha \in [0, 1]\}$ with $\boldsymbol{v}_t^\star = \boldsymbol{f}^\star - \boldsymbol{f}_t = -\boldsymbol{B}_t\nabla\Phi(\boldsymbol{f}_t)$ which sufficiently decreases the objective function:

$$\boldsymbol{f}_{t+1} = \boldsymbol{f}_t + \alpha^\star \boldsymbol{v}_t^\star. \tag{9}$$

To ensure convergence and sufficient progress, the step length $\alpha^\star$ is generally obtained using a "backtracking" algorithm, which consists in gradually decreasing α from an initial value $\alpha^{\text{init}} \leq 1$ until the Wolfe conditions (WCs) are met [29]:

$$\Phi(\boldsymbol{f}_t + \alpha\boldsymbol{v}_t^\star) \leq \Phi(\boldsymbol{f}_t) + \lambda_1 \alpha \nabla\Phi(\boldsymbol{f}_t)^\top \boldsymbol{v}_t^\star \tag{10}$$

$$\|\nabla\Phi(\boldsymbol{f}_t + \alpha\boldsymbol{v}_t^\star)^\top \boldsymbol{v}_t^\star\|_2 \leq \lambda_2 \|\nabla\Phi(\boldsymbol{f}_t)^\top \boldsymbol{v}_t^\star\|_2 \tag{11}$$

where $0 < \lambda_1 < \lambda_2 < 1$ and $\|\cdot\|_2$ is the ℓ^2-norm. In this study, λ_1 and λ_2 were set to 10^{-4} and 0.9, respectively. Since both the objective function and its gradient have to be computed for each new α (as shown in (10)–(11)), extra forward and backward projection operations are required when applying a line search. Note that when $\alpha^\star$ satisfies the WCs and the current estimated $\boldsymbol{B}_t$ is positive-definite, the new estimated L-BFGS matrix $\boldsymbol{B}_{t+1}$ is necessarily positive-definite [12].

2) Boundary Constraints: L-BFGS was extended to L-BFGS-B [15], [16] to be able to handle minimization with box constraints. The search direction is computed by solving the constrained problem corresponding to (7):

$$\boldsymbol{f}^\dagger = \arg\min q_t(\boldsymbol{f}) \quad \text{subject to } \boldsymbol{l} \leq \boldsymbol{f} \leq \boldsymbol{u} \tag{12}$$

where $\boldsymbol{l}$ and $\boldsymbol{u}$ denote the lower and upper bounds of the problem, respectively. In this work, solving (12) was achieved following the method proposed in [15], which utilizes the active constraints defined by the generalized Cauchy point.

We only used a lower boundary constraint $\boldsymbol{l} = \boldsymbol{0}$ to impose the non-negativity constraint of the image reconstruction problem in this study. The line-search is performed in the direction $\boldsymbol{v}_t^\dagger = \boldsymbol{f}^\dagger - \boldsymbol{f}_t$. Similarly to the unbounded case, a backtracking algorithm is used to find a solution $\alpha^\dagger$ that satisfies the WCs. By convexity, the update is guaranteed to satisfy the boundary constraints.

For well-conditioned and small-scale problems, L-BFGS-B is expected to produce a minimizer with fast convergence rate as the approximate $\boldsymbol{H}^{-1}$ is a non-diagonal matrix that takes into account the inter-variable correlation. However, the limited memory approximations that are introduced can lead to low accuracy of the approximate $\boldsymbol{H}^{-1}$ and slow convergence for ill-conditioned or large-scale problems [28].

B. Preconditioned L-BFGS-B (L-BFGS-B-PC)

We propose to circumvent the potential deficiencies of L-BFGS-B via preconditioning. Preconditioning is a general strategy that transforms the problem into a new coordinate system where it is easier to solve (12) [30]. Given $\boldsymbol{f}$ the original estimate, the transformation is described as:

$$\tilde{\boldsymbol{f}} = \boldsymbol{D}\boldsymbol{f} \tag{13}$$

with the preconditioner $\boldsymbol{D}$, the transformation matrix. To deal with the new estimate $\tilde{\boldsymbol{f}}$, the objective function and its derivatives should be transformed accordingly:

$$\begin{aligned} \tilde{\Phi}(\tilde{\boldsymbol{f}}) &= \Phi(\boldsymbol{f}) = \Phi(\boldsymbol{D}^{-1}\tilde{\boldsymbol{f}}) \\ \nabla\tilde{\Phi}(\tilde{\boldsymbol{f}}) &= \boldsymbol{D}^{-1}\nabla\Phi(\boldsymbol{D}^{-1}\tilde{\boldsymbol{f}}) \\ \tilde{\boldsymbol{H}}(\tilde{\boldsymbol{f}}) &= \boldsymbol{D}^{-1}\boldsymbol{H}(\boldsymbol{D}^{-1}\tilde{\boldsymbol{f}})\boldsymbol{D}^{-1} \end{aligned} \tag{14}$$

where $\boldsymbol{H}(\boldsymbol{D}^{-1}\tilde{\boldsymbol{f}})$ and $\tilde{\boldsymbol{H}}(\tilde{\boldsymbol{f}})$ respectively denote the Hessians of Φ and $\tilde{\Phi}$ evaluated at $\boldsymbol{D}^{-1}\tilde{\boldsymbol{f}}$ and $\tilde{\boldsymbol{f}}$. To be able to keep using box-constraints, we propose to use a diagonal preconditioner. Since L-BFGS-B will have to restart the approximation process for constructing $\boldsymbol{B}_t$ every time the preconditioner has been updated, it is essential to use a precomputed (and fixed) preconditioner to prevent constructing $\boldsymbol{B}_t$ with insufficient history iterations. Otherwise, the lack of history information will lead to an unreliable $\boldsymbol{B}_t$ and slow convergence rate.

In our previous study [20], we incorporated the preconditioner introduced in the "precomputed denominator" of relaxed ordered-subsets SPS (OS-SPS) [21] into L-BFGS-B. However, as the preconditioner was calculated with the inverse of the measured data, its performance was sensitive to low counts [23]. The following preconditioner is therefore proposed in this study:

$$\boldsymbol{D} = \operatorname{diag}\left\{\boldsymbol{A}^\top \operatorname{diag}\left\{\frac{\boldsymbol{g}}{(\boldsymbol{A}\boldsymbol{f}^{\text{init}} + \boldsymbol{n})^2}\right\}\boldsymbol{A}\boldsymbol{1} + \beta\nabla^2 R(\boldsymbol{f}^{\text{init}})\boldsymbol{1}\right\}^{\frac{1}{2}} \tag{15}$$

where $\boldsymbol{f}^{\text{init}}$ is the initial guess and $\boldsymbol{1}$ is a vector of ones.

As the preconditioner $\boldsymbol{D}$ is not updated, the overall computational demand of L-BFGS-B-PC is similar to that of L-BFGS-B. Note that the performance of L-BFGS-B-PC will be affected by the initial guess $\boldsymbol{f}^{\text{init}}$. Choosing a better initial guess can therefore improve the convergence rate by starting closer to the solution and also by improving the preconditioner $\boldsymbol{D}$.

Algorithm 1 Pseudo-Code for L-BFGS-B

```
Input: Data $\boldsymbol{g}$, $\Phi$, $\nabla\Phi$, initial $\boldsymbol{f}^{\text{init}}$, $\alpha_0^{\text{init}}$, $\beta$, $\lambda_1$, $\lambda_2$, $m$
Output: Estimated tracer distribution $\boldsymbol{f}$
$\boldsymbol{f}_0 \leftarrow \boldsymbol{f}^{\text{init}}$ ;
$\boldsymbol{d}_0 \leftarrow \nabla\Phi(\boldsymbol{f}_0)$ ;
$\boldsymbol{B} \leftarrow \mathbf{Id}$ ;
for $t = 0, \dots, \text{MaxIter} - 1$ do
    Define
    $q : \boldsymbol{x} \mapsto (\boldsymbol{x} - \boldsymbol{f}_t)^\top \boldsymbol{d}_t + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{f}_t)^\top \boldsymbol{B}^{-1}(\boldsymbol{x} - \boldsymbol{f}_t)$ ;
    $\boldsymbol{f}^\star \leftarrow \arg\min_{\boldsymbol{x} \geq \mathbf{0}} q(\boldsymbol{x})$ ;
    $\boldsymbol{v}^\star \leftarrow \boldsymbol{f}^\star - \boldsymbol{f}_t$ ;
    if $t = 0$ then
        $\alpha^{\text{init}} \leftarrow \alpha_0^{\text{init}}$ ;
    else
        $\alpha^{\text{init}} \leftarrow 1$ ;
    end
    $\alpha^\star \leftarrow \text{WC}(\Phi, \nabla\Phi, \boldsymbol{f}_t, \boldsymbol{v}^\star, \alpha^{\text{init}}, \lambda_1, \lambda_2)$ ;
    $\boldsymbol{f}_{t+1} \leftarrow \boldsymbol{f}_t + \alpha^\star \boldsymbol{v}^\star$ ;
    $\boldsymbol{d}_{t+1} \leftarrow \nabla\Phi(\boldsymbol{f}_{t+1})$ ;
    $m' \leftarrow \min(t + 1, m)$ ;
    $\boldsymbol{B} \leftarrow$
    $\text{ApproxInvHess}\left(\boldsymbol{f}_s, \boldsymbol{d}_s, s \in \{t + 1 - m', \dots, t + 1\}\right)$ ;
end
$\boldsymbol{f} \leftarrow \boldsymbol{f}_{\text{MaxIter}}$ ;
```

C. Implementation

The implementation of L-BFGS-B employed in this study was originally proposed in [15]. A pseudo-code that summarizes the implementation can be found in Algorithm 1.WC refers to the backtracking algorithm to find a step length $\alpha^\star$ which satisfies the WCs (10) and (11), whereas ApproxInvHess refers to the Hessian inverse compact approximation method described in the Appendix. Since it has been observed that a satisfactory approximation of $\boldsymbol{H}^{-1}$ can be obtained based on a few previous iterations [28], a history length $m = 5$ was maintained for constructing $\boldsymbol{B}_t$. To take into account the scale of the variables, at the first iteration, the line search step is initialized by

$$\alpha_0^{\text{init}} = \min\left(\frac{1}{\|\nabla\Phi(\boldsymbol{f}^{\text{init}})\|_2}, 1\right). \tag{16}$$

Although this initialization is fine for certain scales of $\boldsymbol{f}$, for some problems it can lead to suboptimal step length at the first iteration if α_0^{init} is too small. We implemented the algorithm in MATLAB using vendor-provided C functions to calculate the forward and backward projections. In our case, the projector models acquisition on a GE Discovery STE [31].

For the proposed L-BFGS-B-PC, the same L-BFGS-B implementation was used, but with the transformed objective and gradient functions programmed in MATLAB. Since the lack of the scale information of the variables is supplemented by the preconditioner $\boldsymbol{D}$, it is unnecessary for L-BFGS-B-PC to use the suboptimal first step length (16) as L-BFGS-B does. By modifying the initial step length to 1 to match other line searches in the algorithm, L-BFGS-B-PC is able to update the current estimate with a reasonably optimal step length at every iteration. We have verified in initial experiments (not shown) that this modification speeds up the initial line search. Algorithm 2 shows a pseudo-code of the implementation.

Algorithm 2 Pseudo-Code for L-BFGS-B-PC

```
Input: Data $\boldsymbol{g}$, $\Phi$, $\nabla\Phi$, initial $\boldsymbol{f}^{\text{init}}$, $\beta$, $\lambda_1$, $\lambda_2$, $m$
Output: Estimated tracer distribution $\boldsymbol{f}$
$\boldsymbol{f} \leftarrow \boldsymbol{f}^{\text{init}}$ ;
$\boldsymbol{D} \leftarrow \text{diag}\left\{\boldsymbol{A}^\top \text{diag}\left\{\frac{\boldsymbol{g}}{(\boldsymbol{A}\boldsymbol{f}+\boldsymbol{n})^2}\right\}\boldsymbol{A}\mathbf{1} + \beta\nabla^2 R(\boldsymbol{f})\mathbf{1}\right\}^{\frac{1}{2}}$ ;
$\boldsymbol{f} \leftarrow \boldsymbol{D}\boldsymbol{f}$ ;
$\alpha_0^{\text{init}} \leftarrow 1$ ;
Define $\tilde{\Phi} : \boldsymbol{x} \mapsto \Phi(\boldsymbol{D}^{-1}\boldsymbol{x})$ ;
Define $\nabla\tilde{\Phi} : \boldsymbol{x} \mapsto \boldsymbol{D}^{-1}\nabla\Phi(\boldsymbol{D}^{-1}\boldsymbol{x})$ ;
$\boldsymbol{f} \leftarrow \text{L-BFGS-B}(\boldsymbol{g}, \tilde{\Phi}, \nabla\tilde{\Phi}, \boldsymbol{f}, \alpha_0^{\text{init}}, \beta, \lambda_1, \lambda_2, m)$ ;
$\boldsymbol{f} \leftarrow \boldsymbol{D}^{-1}\boldsymbol{f}$ ;
```

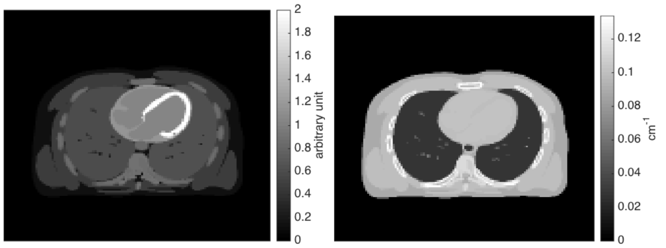

Fig. 1. A slice of the phantom (left) and the corresponding attenuation map (right).

IV. Evaluation

A. Data

The performance of L-BFGS-B and L-BFGS-B-PC was initially evaluated with a digital phantom. To demonstrate the feasibility of practical application, sample reconstructions with three sets of real patient data are also presented.

1) Digital Phantom Simulation: A 3D volume from the XCAT torso phantom [32] was cropped to a 192 × 192 × 47 matrix with voxel size of 3.125 mm. A slice of the phantom and the corresponding attenuation map are shown in Fig 1. This image was forward projected, taking attenuation into account, into 3D sinograms corresponding to data from the GE Discovery STE in 3D acquisition mode. For assessing the noise effects, three data sets with total counts S_{tot} of 52 M, 261 M and 1305 M were generated. Each of them had the same true to background event ratio (TBR) $= 0.74$. We investigated the possible effects from the background by introducing 4 more data sets, which can be divided into two groups. The first group (G1) had the same total counts as the data with $S_{\text{tot}} = 261$ M counts, but had 5 times lower or higher TBR, achieved by adjusting both background S_{bg} and true events S_{true}. As S_{tot} was unchanged, there were less S_{true} in the data with higher S_{bg}. For the other group (G2), we kept S_{true} the same as that in the data with $S_{\text{tot}} = 261$ M counts, but changed S_{bg} by 5 times lower or higher. The total count of the data in G2 after adding the background were $S_{\text{tot}} = 141.2$ M and $S_{\text{tot}} = 860.4$ M,

TABLE I
A SUMMARY OF THE SIMULATED DATA FOR EVALUATING THE INFLUENCE OF THE BACKGROUND

		S_{true}	S_{bg}	S_{tot}
G1	TBR = 0.15	33.7 M	227.3 M	261 M
	TBR = 3.71	205.6 M	55.4 M	261 M
G2	TBR = 0.15	111.2 M	749.2 M	860.4 M
	TBR = 3.71	111.2 M	30 M	141.2 M

respectively. Note that these two groups had identical TBR for the same background level: TBR $= 0.15$ for the high background data and TBR $= 3.71$ for the low background data. Table I shows a summary of the simulated data for evaluating the influence of the background.

2) Patient Data: Data used for this retrospective study included three patient datasets of the thorax acquired on the GE Discovery STE PET/CT scanner. For each study, a cine-CT scan (140 kVp, 60 mA, 4 s duration, 0.5 s rotation period, 0.45 s time between reconstructed images, 9 bed positions, 8 axial slices per bed position) was performed, followed by a PET scan in fully 3D mode. The CT scan was used for the attenuation correction. The acquisition was started 1 hour after the injection of 315 MBq of ^{18}F-FDG and patient consent was collected beforehand. The total counts of the PET data were $S_{\mathrm{tot}} = 181$ M, 255 M and 355 M, respectively. We then used the vendor-provided software to bin the PET data into sinograms and to model the corresponding detection efficiency, attenuation, scatter and randoms.

B. PML Reconstruction

Reconstructed images had $192 \times 192 \times 47$ voxels with voxel size of 3.646 mm. The performance of L-BFGS-B and L-BFGS-B-PC was evaluated using both the quadratic (QP) and log-cosh (LP) penalties. The penalty neighborhood structure was defined as the closest 6 voxels. The scalar ρ in the LP was fixed at 1.8, based on a visual comparison with images from QP, so as to have an apparent edge preserving effect.

C. Initial Image

Initializing reconstruction algorithms with an image closer to the final solution could speed them up, especially for the proposed L-BFGS-B-PC with the preconditioner in (15). To avoid increasing the overall computational cost significantly, we propose to use an initial image reconstructed by ordered-subsets (OS)-type algorithms. In this study, we investigate the use of OS-EM [33] as the algorithm is widely used in practice.

To simplify the problem of finding the best initial image, a two-part study was conducted. In the first part of the study, 8 different numbers of subsets (1, 2, 5, 7, 10, 14, 35 and 70) were employed to speed up the convergence rate. We then fixed the subsets to the limit found in the first part and increased the number of full iterations from one to two to assess if the performance can be improved even further. The reconstruction was then performed by L-BFGS-B-PC initialized with those images described above. The applied penalty function was QP with $\beta = 4$. Note that the initial images were reconstructed without using a penalty function. All initial conditions were evaluated using the digital phantom dataset with $S_{\mathrm{tot}} = 261$ M total counts and TBR $= 0.74$ and the patient data with $S_{\mathrm{tot}} = 355$ M.

D. Analysis

For simulated data, the performance evaluation of L-BFGS-B and L-BFGS-B-PC was conducted in terms of visual comparison, objective function value and a convergence estimate M that measures the distance from the current estimate to the converged image $\boldsymbol{f}^{\mathrm{c}}$. The metric was defined as:

$$M(t) = \sqrt{\frac{1}{\mathrm{N}} \frac{\|\boldsymbol{f}_t - \boldsymbol{f}^{\mathrm{c}}\|_2^2}{(\bar{\boldsymbol{f}}^{\mathrm{c}})^2}} \tag{17}$$

where N is the number of voxels in the volume and $\bar{\boldsymbol{f}}^{\mathrm{c}}$ is the mean value of all voxels in $\boldsymbol{f}^{\mathrm{c}}$. Fast decrease of M indicates fast convergence rate to the solution. For the converged image $\boldsymbol{f}^{\mathrm{c}}$ in (17), we have used the output of SPS [11] at high iteration number, since the convergence of this algorithm has been well-established. To reduce the total computational cost, we used the output of L-BFGS-B-PC with 40 iterations as the initial image for SPS. We then ran SPS for 15000 iterations and investigated the change of visual appearance and objective function values. Since no significant change was observed after 14000 iterations, we chose the image obtained with (L-BFGS-B-PC initialized) SPS at the 15000th iteration as the converged image $\boldsymbol{f}^{\mathrm{c}}$.

An initial evaluation with a visual comparison of a slice of the reconstructed images from both L-BFGS-B and L-BFGS-B-PC at different iterations was used to see if the changes in the convergence rate are relevant. We then performed assessments with respect to penalty type, penalty strength, noise level and TBR to investigate the performance consistency of the algorithms. Quantitative evaluation used plots of both objective function and M values against the total number of projection operations, *i.e.*, the number of projection operations in both the initial OS-EM and L-BFGS-B or L-BFGS-B-PC. Each forward and backward projection of the full set of data was counted separately. We used the number of projection operations instead of the iteration numbers as it represents the computational demand, especially for algorithms involving a line search. Additional computational cost induced by the line search was ignored.

To be able to compare the convergence rate among different datasets, we computed the required number of projection operations and the corresponding iterations for achieving "practical" convergence. The corresponding iteration number was determined by:

$$t_M^{\star} = \min\{t : M(t) \leq 0.01\}. \tag{18}$$

We also demonstrated the performance of the algorithms with patient data.

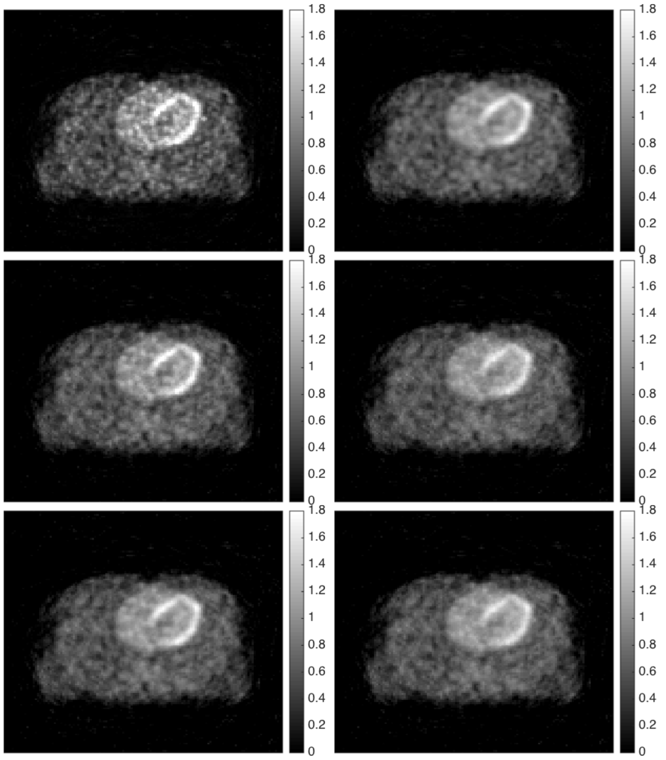

Fig. 2. A slice of the data with 261 M counts (TBR = 0.74) reconstructed by L-BFGS-B (left column) and L-BFGS-B-PC (right column) at the 5th (first row), 10th (second row) and 15th (third row) iteration.

V. Results

A. Initial Investigation

Fig. 2 shows reconstructed images of the XCAT data with $S_{\mathrm{tot}} = 261$ M counts and TBR $= 0.74$ at the 5th, 10th and 15th iteration for L-BFGS-B and L-BFGS-B-PC. Both algorithms are initialized by the best initial image found in section V-B. The reconstructions were performed with QP and $\beta = 20$. Comparing images at the same iteration, we found those from L-BFGS-B-PC represent better contrast and object delineation than images reconstructed by the other algorithm.

Images for L-BFGS-B and L-BFGS-B-PC at iterations that achieve convergence of M values (18) are shown in Fig. 3, with the converged image from SPS for comparison. Profiles along the central row of each image are also provided. As shown in the figure, both algorithms are able to converge visually to the same image and profile as SPS does.

An example comparison of the convergence rate of M values for L-BFGS-B and L-BFGS-B-PC with the modified line search is given in Fig. 4. Results for SPS are also provided. As shown in the plot, both L-BFGS-B and L-BFGS-B-PC achieved several times faster convergence rate than SPS. Also, the proposed L-BFGS-B-PC shows the ability to converge rapidly compared to L-BFGS-B. Although only images from one simulation condition are provided, similar behavior was observed for all studied data and reconstruction configurations. More comparison results for these algorithms can be found in our previous study [23].

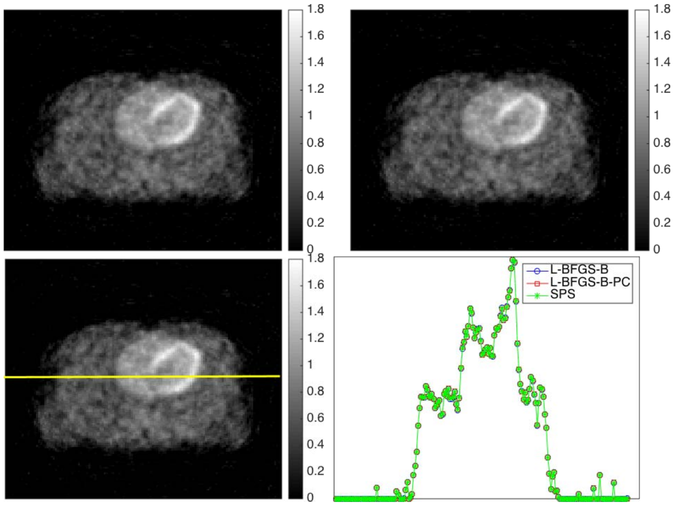

Fig. 3. A slice of images that achieves convergence of M values for L-BFGS-B (at the 44th iteration) (top-left) and L-BFGS-B-PC (at the 24th iteration) (top-right). The converged image from SPS is also shown for comparison (bottom-left). Profiles along the central row of all images are also provided (bottom-right).

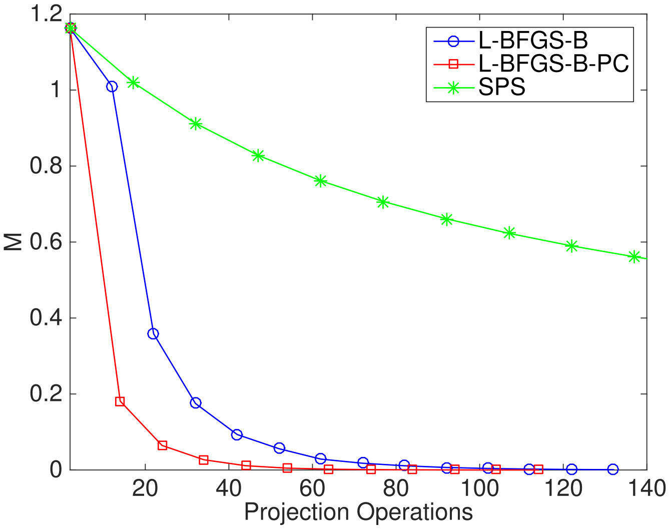

Fig. 4. A comparison of the convergence rate of M values for SPS, L-BFGS-B and L-BFGS-B-PC with respect to the total projection operations.

B. Initial Image

The convergence rate was evaluated by plotting the objective function value against the total number of projection operations. As shown in Fig. 5 top, the convergence rate was improved as the number of subsets was increased. The convergence trend for 70 subsets (there were only 4 projections in one subset) was quite different from the others. Therefore, we chose 35 as the highest number of subsets and increased the full iteration number. Based on the results in Fig. 5 bottom, the performance was not improved any further after one full iteration. Although not shown here, similar results were observed with the patient data. All reconstructions were therefore initialized by 1 iteration of OS-EM with 35 subsets. Note that we did not plot results from the initial point to improve

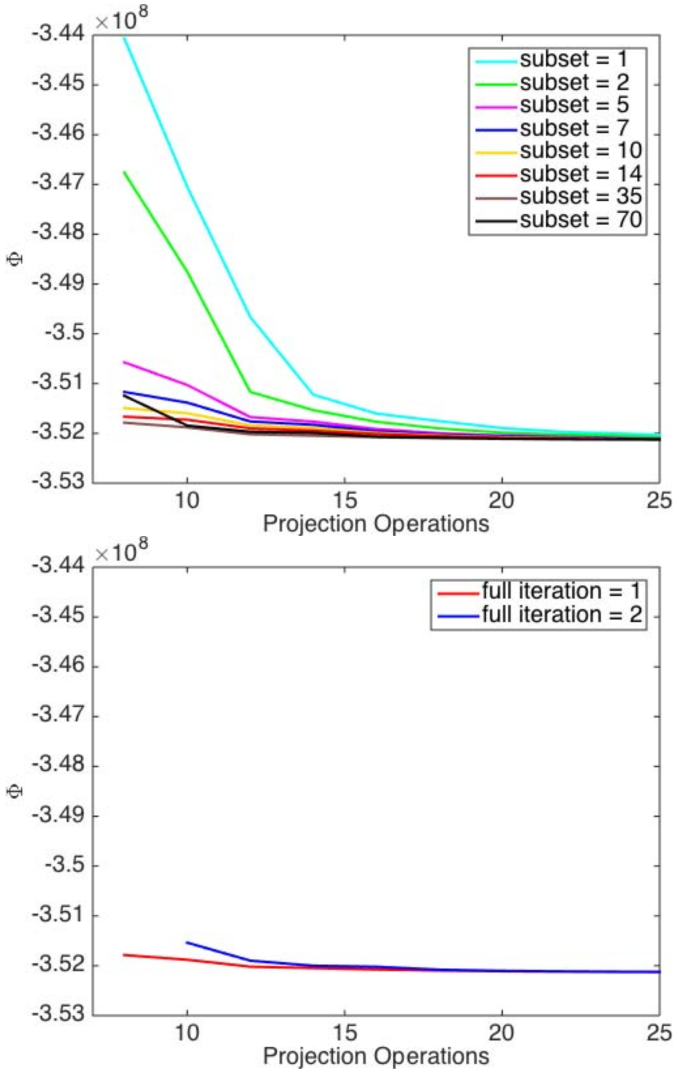

Fig. 5. The objective function values plotted against projection operations for LBFGS-B-PC initialized by one full iteration of OS-EM with various subsets (top) and 2 different full iterations of OS-EM with 35 subsets (bottom).

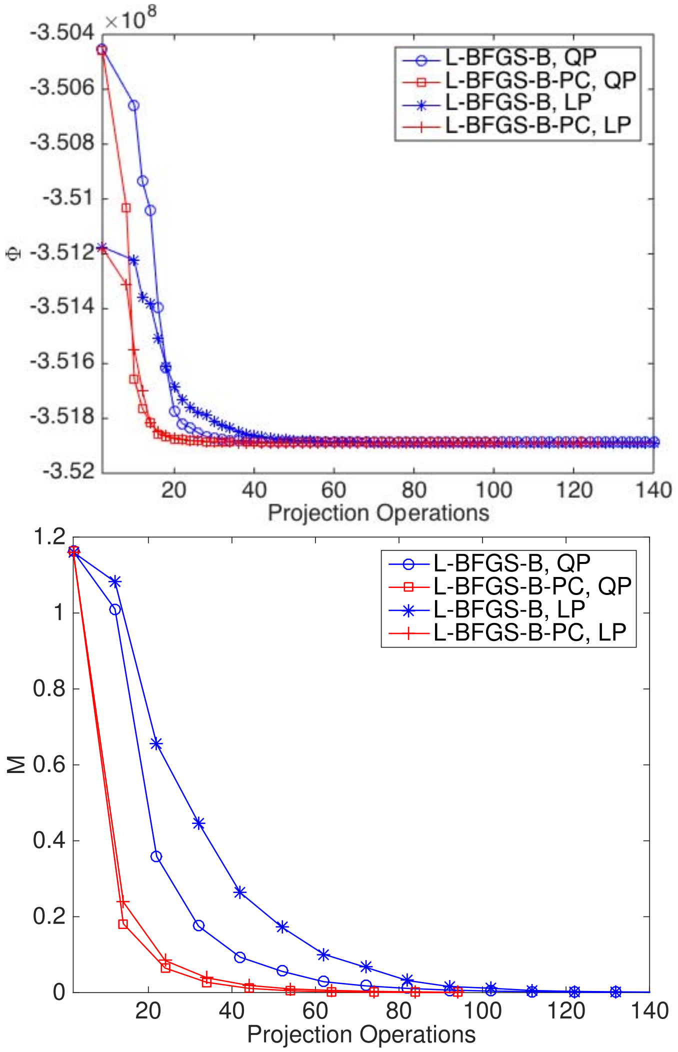

Fig. 6. Objective function values (top) and M values (bottom) plotted against the total projection operations.

clarity. The first point of each line represents the objective function value after the first iteration and the corresponding total projection operation is the required projection operations for constructing the initial image plus that for completing the first iteration of L-BFGS-B-PC.

C. Convergence Rate

The objective function values plotted against the total projection operations are shown in Fig. 6 top. We used results from the same dataset and reconstruction configuration as in the visual comparison section as an example. Both algorithms tend to converge to the same value but with different speeds. By introducing a preconditioner, L-BFGS-B-PC converged rapidly in terms of the objective function value. Fig. 6 bottom is the corresponding M values plotted against projection operations. Similar to the plot of the objective function values, L-BFGS-B-PC achieves superior convergence rate of M value to L-BFGS-B. Moreover, the difference in performance for the two penalty types is extremely small for the proposed algorithm. Consistent results are obtained for other simulated conditions.

D. Convergence Dependence on Different Factors

Simulated data with $S_{\text{tot}} = 52$ M, 261 M and 1305 M, representing high, medium and low noise level, were reconstructed by L-BFGS-B and L-BFGS-B-PC with both $\beta = 4$ and $\beta = 20$ to investigate the effect of noise level and penalty strength. The smoothing QP or edge preserving LP penalty functions were used for evaluating the performance dependence on the penalty type. For each condition, Table II lists the required number of projection operations for achieving convergence according to (18). Values for simulation conditions with noise level from high to low are shown from left to right and separated by a slash. We also listed in parentheses the corresponding number of iterations. Except for the low noise data reconstructed using L-BFGS-B-PC with LP and $\beta = 20$, both algorithms generally required more operations (or iterations) to satisfy the convergence criterion as the noise level was increased or the

TABLE II

THE REQUIRED NUMBER OF PROJECTION OPERATIONS AND ITERATIONS FOR ACHIEVING CONVERGENCE OF *M* VALUES FOR DIFFERENT PENALTY TYPES, PENALTY STRENGTHS AND NOISE LEVELS

		L-BFGS-B	L-BFGS-B-PC
QP	$\beta = 4$	162 / 132 / 82[1] (79 / 64 / 40)[2]	94 / 64 / 44 (45 / 29 / 20)
	$\beta = 20$	122 / 92 / 72 (59 / 44 / 35)	64 / 54 / 44 (30 / 24 / 20)
LP	$\beta = 4$	552 / 182 / 132 (274 / 89 / 65)	94 / 84 / 64 (45 / 40 / 30)
	$\beta = 20$	182 / 112 / 92 (89 / 54 / 44)	74 / 54 / 84 (35 / 25 / 39)

[1] Values listed from left to right and separated by a slash are the required numbers of projection operations for problems with noise level from high to low.

[2] Values listed in parentheses are the corresponding number of iterations.

penalty strength was decreased. Note that reconstructing with LP led to a slower convergence rate than when using QP (with the same β). The possible cause of the exception is discussed in section VI.

The data simulating different background levels in both groups of fixed S_{tot} and fixed number of S_{true} were used to study the influence of the background on the convergence rate. The results were compared with those from the data with $S_{\text{tot}} = 261$ M counts and TBR $= 0.74$ ($S_{\text{true}} = 111.2$ M and $S_{\text{bg}} = 149.8$ M). Since the dependence on penalty type and strength were included above, the data were reconstructed with only QP and $\beta = 4$ for both algorithms. We evaluated the convergence rate by plotting M values against the total projection operations (Fig. 7) and by listing the required number of projection operations to reach the convergence of M values (Table III). The former shows the convergence rate in early iterations while the latter quantifies this at late iterations. For data with the same S_{tot}, the higher the TBR value (*i.e.*, the more true events) the faster the convergence rate in early iterations is observed (Fig. 7 top). However, an opposite trend is obtained when S_{tot} is increased with the background level. The presence of the background helps the convergence rate in early iterations when the same number of S_{true} are collected (Fig. 7 bottom).

Considering the convergence rate at later iterations, we found that data with the same S_{tot} can reach the criterion (18) at almost the same iteration, regardless of the change in the background level. For data with a fixed number of S_{true} but increasing TBR, more iterations are needed to achieve the convergence of M values (Table III). Despite the observed dependence on various factors, the proposed L-BFGS-B-PC shows a relatively consistent performance and outperforms L-BFGS-B in all cases.

E. Demonstration With Patient Data

The patient data were reconstructed by both algorithms with QP and a fixed $\beta = 20$. A coronal view of one patient dataset from each algorithm at the iteration that achieves criterion (18) are shown in Fig. 8 as an example. Profiles along the central slice of both images are also provided. As in the simulation study, the algorithms are able to converge to visually identical images. This was also the case for the other two patient data sets (not shown). Fig. 9 shows the M values plotted against the total projection operations for each algorithm. Faster initial convergence is achieved for data with higher S_{tot}, which is similar to what was observed in Table II. The required projection operations for achieving the convergence of M values are listed in Table IV. Based on the results in Fig. 9 and Table IV, L-BFGS-B-PC shows faster convergence rate than L-BFGS-B in all cases and its performance is much less sensitive to noise level.

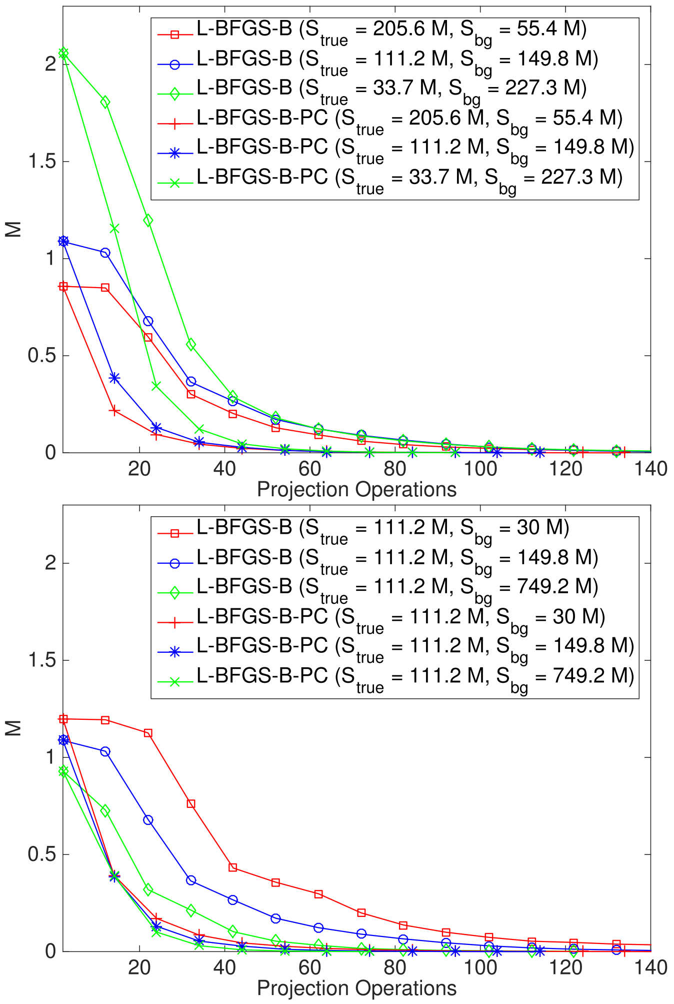

Fig. 7. *M* values plotted against the total projection operations for data with a fixed number of S_{tot} (top) and data with a fixed number of S_{true} (bottom) but different background levels.

VI. DISCUSSION

We have demonstrated the feasibility of using L-BFGS-B and L-BFGS-B-PC in PML reconstruction problems in ET. Both L-BFGS-B and L-BFGS-B-PC are able to converge to virtually identical solutions as SPS (Fig. 3) but with different speed. For the evaluation of the computational demand, we used the total projection operations instead of the computation time because the algorithms were not implemented using the same programming language. For example, L-BFGS-B was

TABLE III
THE REQUIRED NUMBERS OF PROJECTION OPERATIONS AND ITERATIONS FOR ACHIEVING CONVERGENCE OF M VALUES FOR DATA WITH DIFFERENT S_{TOT}, S_{TRUE} AND BACKGROUND LEVELS

	L-BFGS-B	L-BFGS-B-PC
S_{tot} = 261 M, TBR = 0.74	132 (64)[1]	64 (29)
S_{tot} = 261 M, TBR = 3.71	132 (65)	64 (31)
S_{tot} = 261 M, TBR = 0.15	142 (70)	64 (31)
S_{tot} = 141.2 M, TBR = 3.71	272 (135)	84 (39)
S_{tot} = 860.4 M, TBR = 0.15	92 (46)	44 (21)

[1] The required numbers of projection operations and the corresponding number of iterations for each reconstruction are listed together with the latter in parentheses.

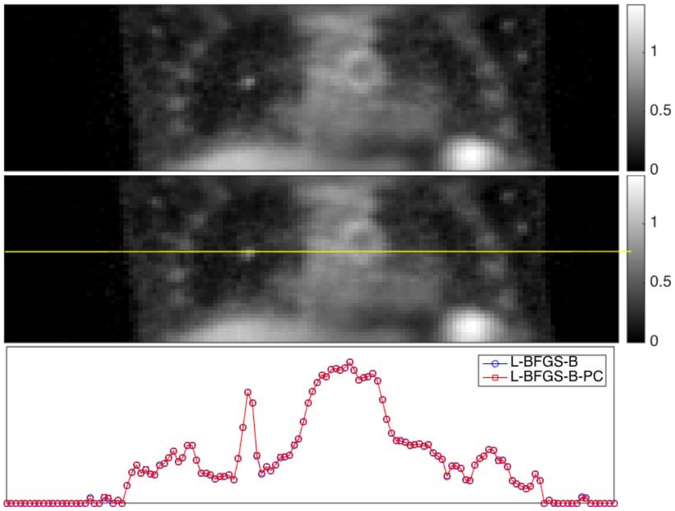

Fig. 8. A coronal view of images for L-BFGS-B at the 260st iteration (top) and L-BFGS-B-PC at the 35th iteration (median) from one patient data. Profiles along the central slice of both images are also provided (bottom).

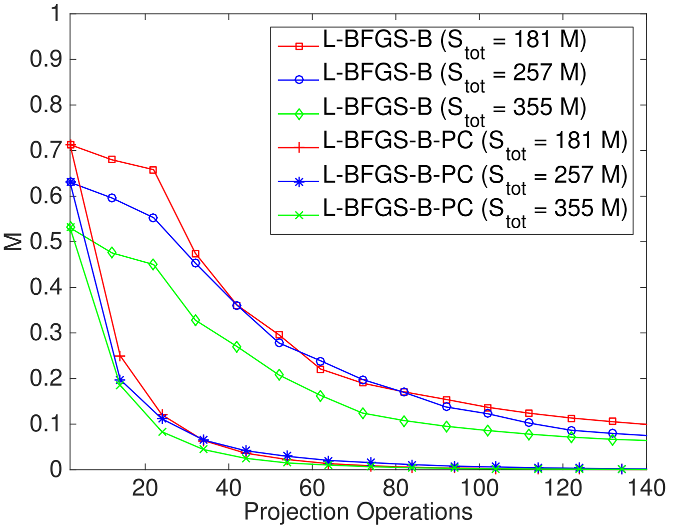

Fig. 9. The M values plotted against the total projection operations for all patient data.

TABLE IV
THE REQUIRED PROJECTION AND ITERATION NUMBERS FOR ACHIEVING CONVERGENCE OF M VALUES FOR 3 PATIENT DATA SETS

	L-BFGS-B	L-BFGS-B-PC
S_{tot} = 181 M	662 (330)	74 (36)
S_{tot} = 257 M	752 (375)	94 (45)
S_{tot} = 355 M	522 (260)	74 (35)

implemented in a combination of C, Fortran and MATLAB while SPS was implemented in MATLAB but using the vendor provided projectors programmed in C. Therefore, except for the visual comparison that requires results at certain iterations (Fig. 2, 3 and 8), we used plots and tables based on projection operations to compare the computational demand between different algorithms. In terms of memory demand, however, both L-BFGS-B and L-BFGS-B-PC require more memory for storing the correction matrices used to represent $\boldsymbol{B}_t$ comparing to SPS. The required extra memory is approximately twice of the product of the total number of voxels J and the maintained history length m (see Appendix for more information). As a precomputed preconditioner has to be stored as well for the proposed L-BFGS-B-PC, it uses slightly more memory than L-BFGS-B.

To quantify the convergence rate, we introduced an image-based metric M measuring the distance from the current estimate to the expected solution. Comparing the top plot of Fig. 6 to the bottom one, we found that both L-BFGS-B and L-BFGS-B-PC required a higher number of projection operations to reach a stable M value than to reach a stable objective function value. We have therefore concentrated on the convergence of M values in this discussion. Moreover, since we observed that the convergence rate of M values for SPS is much slower than that for both L-BFGS-B and L-BFGS-B-PC (Fig. 4 and our previous work [20]), we have excluded SPS in further comparison or discussion.

In our previous work [20], we also compared the convergence rate of L-BFGS-B and the proposed L-BFGS-B-PC with OSL-EM [8] and relaxed SPS [21]. We found that both L-BFGS based algorithms were able to converge over 10 times faster than the others in terms of objective function value and regional recovery ratio in early iterations. The convergence rate of OSL-EM and relaxed SPS can be further improved by using ordered subsets. However, the former algorithm will then suffer from the limit cycle problem while the performance of the latter will depend on the relaxation parameter. Both issues make the comparison of the convergence rate difficult, especially at late iterations. Therefore, we did not include OSL-EM and relaxed SPS for comparison in this study.

In studying the dependence of the convergence rate of L-BFGS-B and L-BFGS-B-PC on various factors, we observed that faster convergence rate was achieved generally with a smoothing prior, strong penalty strength and low noise level data for both algorithms (Table II). In terms of convergence rate of both M values and the objective function values, the proposed L-BFGS-B-PC outperformed L-BFGS-B for all datasets that have been evaluated. In particular, L-BFGS-B-PC achieved convergence within 100 projection operations for all simulations, even for the noisy data set (*i.e.,* the simulated data with $S_{\mathrm{tot}} = 52$ M). The results suggest that the proposed algorithm can even be used in cases where the noise level is high, such as in gated or dynamic studies.

Based on Fig. 7 top, we found that the change in background level for data with the same S_{tot} can affect the convergence rate in early iterations. From the plot, this can be at least partially explained by the fact that the initial OS-EM image was further away from the final solution for a higher background. Other algorithms, less sensitive to true to background ratio for initialization, might decrease this effect. Despite the performance dependence on those factors, at later iterations, the proposed L-BFGS-B-PC is more consistent compared to L-BFGS-B (Table II and Table III).

For the patient data study, the data set with the highest S_{tot} achieved the fastest convergence rate for both L-BFGS-B and L-BFGS-B-PC at late iterations, which is consistent with what has been observed from the simulation study. However, the slowest convergence rate of M values was observed for the data with medium S_{tot}. Since the performance could be affected by many factors, such as patient size and scatter fraction, a more comprehensive evaluation with more patients would be necessary.

Recall that the number of projection operations includes both the forward and backward projections in the combined OS-EM and L-BFGS, and the line search. As shown in the tables, we found that for both algorithms the number of projection operations for achieving the convergence criterion is only slightly larger than twice the number of iterations. This means that the line search subroutine did not involve many projections and so required minimum computational burden. In other words, the initial step length satisfied the WCs (10) and (11) for almost every iteration. As mentioned in section III-C, both algorithms initialize the line search with a step length of 1 after the first iteration. With this step length, the algorithms make a direct approach from the current estimate to the local solution of (12) as described in section III-A.2 with $\boldsymbol{v}_t^\dagger = \boldsymbol{f}^\dagger - \boldsymbol{f}_t$. The backtracking of the embedded line search takes place only when the algorithm is about to converge. To find a smaller step length, a certain decreasing pattern predefined by the backtracking algorithm is considered. However, depending on the adopted decrease scheme, the backtracking might not be able to find the step length that minimizes the objective function for the reconstruction algorithm at the current estimate. We suspect that this is the cause of the unexpected slow convergence rate observed in Table III for the last entry (*i.e.*, the required projection operations for achieving the convergence of M values for the low noise data reconstructed by L-BFGS-B-PC with LP and $\beta = 20$). Further optimization of the line search is beyond the scope of this study.

The primary motivation of incorporating a preconditioner into L-BFGS-B is to have an initial estimate of the second derivative associated with the problem. By utilizing the extra information from the start, L-BFGS-B-PC is able to solve the reconstruction problem rapidly and shows consistent performance for different data conditions and reconstruction configurations. Although the current paper concentrated on L-BFGS-B, the proposed strategy could be applied to other algorithms as well.

In this paper, we have used the preconditioner (15). Additional information will be provided by $\boldsymbol{B}_t$ after a few iterations and the influence of the preconditioner will become less significant. This implies that the preconditioner does not need to be a precise approximation of the square root of the Hessian. Therefore, the algorithm should be able to benefit from other fixed diagonal approximations of the square root of the Hessian. For example, by expressing ML-EM in a gradient descent form, a diagonal matrix with elements equal to a normalized version of the current estimate was obtained in [34]. This was used as motivation for using this diagonal matrix as a preconditioner to improve the convergence rate of a conjugate gradient algorithm [35]. In that paper, the preconditioner was updated at each iteration. However, in order for L-BFGS-B to benefit from the previous iterations when constructing $\boldsymbol{B}_t$, we can replace the current estimate by the initial image as for the proposed preconditioner so that the preconditioner becomes pre-computable.

In this study, we used QP and LP as the penalty functions since both are convex and twice differentiable. This supports the use of L-BFGS-B which approximates the local estimate of the second derivative of a function by differences of first derivatives. In the case where the function being minimized is differentiable but not twice differentiable at some point (e.g. the Huber functional [36]), it is likely that the L-BFGS-B algorithm will have difficulty. Investigating additional priors, however, is beyond the scope of this paper.

VII. Conclusion

We have investigated the performance of L-BFGS-B for penalized reconstruction problems in ET with simulated and real patient data. Its convergence rate can be considerably improved by introducing a diagonally-scaled preconditioner (L-BFGS-B-PC) combined with good initialization. Since the proposed preconditioner can be precomputed, the overall computational demand of L-BFGS-B-PC is similar to that of L-BFGS-B. In addition to showing faster convergence rate than L-BFGS-B, the performance of L-BFGS-B-PC, in terms of the objective function value and the image-based metric M, is less sensitive to penalty type, penalty strength, data noise level and background level. These encouraging results indicate the potential usefulness of L-BFGS-B-PC for achieving high quantitative accuracy with acceptable reconstruction time.

Appendix
Construction of the Approximation of the Inverse of the Hessian Using a Pair of Correction Matrices

This section describes the ApproxInvHess step in Algorithm 1. At every iteration t, the corresponding correction matrices consisting of gradient information in the last m iterations are expressed as follows:

$$\boldsymbol{S}_t = [\boldsymbol{s}_{t-m}, \ldots, \boldsymbol{s}_{t-1}], \quad \boldsymbol{Y}_t = [\boldsymbol{y}_{t-m}, \ldots, \boldsymbol{y}_{t-1}] \qquad (19)$$

where $\boldsymbol{s}_t = \boldsymbol{f}_{t+1} - \boldsymbol{f}_t$ and $\boldsymbol{y}_t = \nabla\Phi(\boldsymbol{f}_{t+1}) - \nabla\Phi(\boldsymbol{f}_t)$. These matrices can be used to find the 2nd order behavior of the objective function and therefore to calculate approximations of the Hessian. Based on the compact representations described in [28], the approximation of $\boldsymbol{H}^{-1}$ at iteration t can be written as follows:

$$\boldsymbol{B}_t \equiv \frac{1}{Q}\boldsymbol{I} + \bar{\boldsymbol{W}}_t\bar{\boldsymbol{M}}_t\bar{\boldsymbol{W}}_t^\top \tag{20}$$

where

$$\begin{aligned}
\bar{\boldsymbol{W}}_t &\equiv \begin{bmatrix} \frac{1}{Q}\boldsymbol{Y}_t & \boldsymbol{S}_t \end{bmatrix}, \\
\bar{\boldsymbol{M}}_t &\equiv \begin{bmatrix} \boldsymbol{0} & -\boldsymbol{R}_t^{-1} \\ -\boldsymbol{R}_t^{-\top} & \boldsymbol{R}_t^{-\top}(\boldsymbol{V}_t + \frac{1}{Q}\boldsymbol{Y}_t^\top\boldsymbol{Y}_t\boldsymbol{R}_t^{-1}) \end{bmatrix}, \\
[\boldsymbol{R}_t]_{kl} &= \begin{cases} \boldsymbol{s}_{t-m-1+k}^\top \boldsymbol{y}_{t-m-1+l} & \text{if } k \le l \\ \boldsymbol{0} & \text{otherwise} \end{cases}
\end{aligned}$$

with $\boldsymbol{V}_t = \operatorname{diag}\{\boldsymbol{s}_{t-m}^\top\boldsymbol{y}_{t-m}, \ldots, \boldsymbol{s}_{t-1}^\top\boldsymbol{y}_{t-1}\}$, $k, l = 1, \ldots, m$ and Q is a constant [15]. The representation of $\boldsymbol{B}_t$ is efficient in terms of memory and computation time as $\bar{\boldsymbol{W}}_t$ is a $J \times 2m$ matrix and $\bar{\boldsymbol{M}}_t$ is $2m \times 2m$, where J is the number of voxels and $m = 5$ in this study. In practice, the algorithm does not compute and store $\boldsymbol{B}_t$ directly. Instead, it uses the correction matrices so that the product $\boldsymbol{B}_t\nabla\Phi(\boldsymbol{f}_t)$ can be calculated efficiently by applying the unrolling technique described in [28].

To initialize the construction of $\boldsymbol{B}_1$, the current implementation performs gradient descent at the first iteration to find the first pair of correction vectors, $\boldsymbol{S}_1 = [\boldsymbol{s}_0]$ and $\boldsymbol{Y}_1 = [\boldsymbol{y}_0]$. For iteration $t < m$, the corresponding $\boldsymbol{B}_t$ is calculated with only t pairs of gradient information.

References

[1] D. L. Bailey and K. P. Willowson, "An evidence-based review of quantitative SPECT imaging and potential clinical applications," *J. Nucl. Med.*, vol. 54, no. 1, pp. 83–89, 2013.

[2] J. Y. Ngeow *et al.*, "High SUV uptake on FDG–PET/CT predicts for an aggressive B-cell lymphoma in a prospective study of primary FDG–PET/CT staging in lymphoma," *Ann. Oncol.*, vol. 20, no. 9, pp. 1543–1547, 2009.

[3] B. Bai, J. Bading, and P. S. Conti, "Tumor quantification in clinical positron emission tomography," *Theranostics*, vol. 3, no. 10, pp. 787–801, 2013.

[4] V. P. Hart, II, "The application of tomographic reconstruction techniques to ill-conditioned inverse problems in atmospheric science and biomedical imaging," Ph.D. dissertation, Dept. Phys., Utah State Univ., Logan, UT, USA, 2012.

[5] J. Dutta, S. Ahn, and Q. Li, "Quantitative statistical methods for image quality assessment," *Theranostics*, vol. 3, no. 10, pp. 741–756, 2013.

[6] J. M. Ollinger and J. A. Fessler, "Positron-emission tomography," *IEEE Signal Process. Mag.*, vol. 14, no. 1, pp. 43–55, Jan. 1997.

[7] A. Alessio and P. Kinahan, *PET Image Reconstruction*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2006.

[8] P. J. Green, "On use of the EM for penalized likelihood estimation," *J. Roy. Stat. Soc. B, (Methodol.)*, vol. 52, no. 3, pp. 443–452, 1990.

[9] A. R. De Pierro and M. E. B. Yamagishi, "Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography," *IEEE Trans. Med. Imag.*, vol. 20, no. 4, pp. 280–288, Apr. 2001.

[10] A. R. De Pierro, "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.

[11] H. Erdoğan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, vol. 44, no. 1, pp. 2835–2851, 1999.

[12] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed. Springer, 2006, pp. 134–141.

[13] H. Matthies and G. Strang, "The solution of nonlinear finite element equations," *SIAM J. Sci. Comput.*, vol. 14, no. 11, pp. 1613–1626, 1979.

[14] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Math. Comput.*, vol. 35, no. 151, pp. 773–782, 1980.

[15] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995.

[16] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997.

[17] G. Andrew and J. Gao, "Scalable training of L^1-regularized log-linear models," in *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 33–40.

[18] M. J. Ehrhardt *et al.*, "PET reconstruction with an anatomical MRI prior using parallel level sets," *IEEE Trans. Med. Imag.*, vol. 35, no. 9, pp. 2189–2199, Sep. 2016.

[19] S. Somayajula, C. Panagiotou, A. Rangarajan, Q. Li, S. R. Arridge, and R. M. Leahy, "PET image reconstruction using information theoretic anatomical priors," *IEEE Trans. Med. Imag.*, vol. 30, no. 3, pp. 49–537, Mar. 2011.

[20] Y.-J. Tsai, A. Bousse, M. J. Ehrhardt, B. F. Hutton, S. Arridge, and K. Thielemans, "Performance evaluation of MAP algorithms with different penalties, object geometries and noise levels," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. Rec.*, Oct. 2015, pp. 1–3.

[21] S. Ahn and J. A. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," *IEEE Trans. Med. Imag.*, vol. 22, no. 5, pp. 613–626, May 2003.

[22] M. S. Kaplan, D. R. Haynor, and H. Vija, "A differential attenuation method for simultaneous estimation of SPECT activity and attenuation distributions," *IEEE Trans. Nucl. Sci.*, vol. 46, no. 3, pp. 41–535, Jun. 1999.

[23] Y.-J. Tsai *et al.*, "Performance improvement and validation of a new MAP reconstruction algorithm," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. Rec.*, Oct. 2016, pp. 1–3.

[24] K. M. Hanson, "Introduction to Bayesian image analysis," *Proc. SPIE*, pp. 716–731, Sep. 1993.

[25] K. Vunckx *et al.*, "Evaluation of three MRI-based anatomical priors for quantitative PET brain imaging," *IEEE Trans. Med. Imag.*, vol. 31, no. 3, pp. 599–612, Mar. 2012.

[26] B. Bai, Q. Li, and R. M. Leahy, "MR guided PET image reconstruction," *Semin. Nucl. Med.*, vol. 43, no. 1, pp. 30–44, 2013.

[27] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization," Dept. EECS, Northwestern Univ., Evanston, IL, USA, Tech. Rep. NAM12, 1995.

[28] R. H. Byrd, J. Nocedal, and R. B. Schnabel, "Representations of quasi-Newton matrices and their use in limited memory methods," *Math. Programm.*, vol. 63, no. 1, pp. 129–156, 1994.

[29] J. J. Moré and D. J. Thuente, "On line search algorithms with guarenteed sufficient decrease," *ACM Trans. Math. Softw.*, vol. 20, no. 3, pp. 286–307, 1994.

[30] T. Allahviranloo, R. G. Moghaddam, and M. Afshar, "Comparison theorem with modified Gauss-Seidel and modified Jacobi methods by M-matrix," *J. Interpolation Approx. Sci. Comput.*, vol. 2012, Sep. 2012, Art. no. jiasc-00017.

[31] M. Teräs, T. Tolvanen, J. J. Johansson, J. J. Williams, and J. Knuuti, "Performance of the new generation of whole-body PET/CT scanners: Discovery STE and discovery VCT," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 34, no. 10, pp. 92–1683, 2007.

[32] W. P. Segars, G. Sturgeon, S. Mendonca, J. Grimes, and B. M. W. Tsui, "4D XCAT phantom for multimodality imaging research," *Med. Phys.*, vol. 37, no. 9, pp. 4902–4915, 2010.

[33] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 601–609, Dec. 1994.

[34] 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, Jun. 1987.

[35] 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. 4, pp. 687–701, Dec. 1994.

[36] J. A. Fessler, E. P. Ficaro, N. H. Clinthorne, and K. Lange, "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 166–175, Apr. 1997.