Fast X-Ray CT Image Reconstruction Using a Linearized Augmented Lagrangian Method with Ordered Subsets

Hung Nien, *Student Member, IEEE*, and Jeffrey A. Fessler, *Fellow, IEEE*

***Abstract*—Augmented Lagrangian (AL) methods for solving convex optimization problems with linear constraints are attractive for imaging applications with composite cost functions due to the empirical fast convergence rate under weak conditions. However, for problems such as X-ray computed tomography (CT) image reconstruction, where the inner least-squares problem is challenging and requires iterations, AL methods can be slow. This paper focuses on solving regularized (weighted) least-squares problems using a linearized variant of AL methods that replaces the quadratic AL penalty term in the scaled augmented Lagrangian with its separable quadratic surrogate (SQS) function, leading to a simpler ordered-subsets (OS) accelerable splitting-based algorithm, OS-LALM. To further accelerate the proposed algorithm, we use a second-order recursive system analysis to design a deterministic downward continuation approach that avoids tedious parameter tuning and provides fast convergence. Experimental results show that the proposed algorithm significantly accelerates the convergence of X-ray CT image reconstruction with negligible overhead and can reduce OS artifacts when using many subsets.**

***Index Terms*—Statistical image reconstruction, computed tomography, ordered subsets, augmented Lagrangian.**

I. Introduction

STATISTICAL methods for image reconstruction have been explored extensively for computed tomography (CT) due to the potential of acquiring CT scans with lower X-ray dose while maintaining image quality. However, the much longer computation time of statistical methods still restrains their applicability in practice. To accelerate statistical methods, many optimization techniques have been investigated. Augmented Lagrangian (AL) methods (including the alternating direction variants) [1–4] are powerful techniques for solving regularized inverse problems using variable splitting. For example, in total-variation (TV) denoising and compressed sensing (CS) problems, AL methods can separate non-smooth ℓ_1 regularization terms by introducing auxiliary variables, yielding simple penalized least-squares inner problems that are solved efficiently using the fast Fourier transform (FFT) algorithm and proximal mappings such as the soft-thresholding for the ℓ_1-norm [5, 6]. However, in applications like X-ray CT image reconstruction, the inner least-squares problem is challenging due to the highly shift-variant Hessian caused by the huge dynamic range of the statistical weighting. To solve this problem, Ramani *et al.* [7] introduced an additional variable that separates the shift-variant and approximately shift-invariant components of the statistically weighted quadratic data-fitting term, leading to a better-conditioned inner least-squares problem that was solved efficiently using the preconditioned conjugate gradient (PCG) method with an appropriate circulant preconditioner. Experimental results showed significant acceleration in 2D CT [7]; however, in 3D CT with cone-beam geometry, it is more difficult to construct a good preconditioner for the inner least-squares problem, and the method in [7] has yet to achieve the same acceleration as in 2D CT. Furthermore, even when a good preconditioner can be found, the iterative PCG solver requires several forward/back-projection operations per outer iteration, which is very time-consuming in 3D CT, significantly reducing the number of outer-loop image updates one can perform within a given reconstruction time.

The ordered-subsets (OS) algorithm [8] is a first-order method with a diagonal preconditioner that uses somewhat conservative step sizes but is easily applicable to 3D CT. By grouping the projections into M ordered subsets that satisfy the “subset balance condition” and updating the image incrementally using the M subset gradients, OS algorithms effectively perform M times as many image updates per outer iteration as the standard gradient descent method, leading to M times acceleration in early iterations. We can interpret the OS algorithm and its variants as incremental gradient methods [9]; when the subset is chosen randomly with some constraints so that the subset gradient is unbiased and with finite variance, they can also be called stochastic gradient methods [10]. Recently, OS variants [11, 12] of the fast gradient method [13–15] demonstrated dramatic acceleration (about M^2 times in early iterations) over their one-subset counterparts. However, when M increases, fast OS algorithms seem to have “larger” limit cycles and exhibit artifacts in the reconstructed images. This problem is also studied in the machine learning literature. Devolder showed that the error accumulation in fast gradient methods is inevitable when an inexact oracle is used, but it can be reduced by using relaxed momentum, i.e., a growing diagonal majorizer (or equivalently, a diminishing step size), at the cost of slower convergence rate [16]. Schmidt *et al.* also showed that an accelerated proximal gradient method is more sensitive to errors in the gradient and proximal mapping calculations [17].

OS-based algorithms, including the standard one and its fast variants, are not convergent in general (unless relaxation [18] or incremental majorization [19] is used, unsurprisingly, at

This work is supported in part by National Institutes of Health (NIH) grants R01-HL-098686 and U01-EB-18753 and by an equipment donation from Intel Corporation. Hung Nien and Jeffrey A. Fessler are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, USA (e-mail: {hungnien,fessler}@umich.edu).

the cost of slower convergence rate) and possibly introduce noise-like artifacts. Nevertheless, the effective M-times image updates using OS is still very promising for AL methods. As an example of combining OS (or stochastic gradients) with AL methods, Ouyang *et al.* [20] proposed a stochastic setting for the alternating direction method of multipliers (ADMM) [4, 6] that introduces an auxiliary variable for the regularization term and majorizes the smooth data-fitting term such as the logistic loss in the scaled augmented Lagrangian using a diagonal majorizer with stochastic gradients. For every stochastic ADMM iteration, only part of the data is visited (for evaluating the gradient of a subset of the data). This substantially reduces the cost per stochastic ADMM iteration, and one can run more stochastic ADMM iterations in a given reconstruction time. However, stochastic ADMM simply combines the stochastic gradient method and ADMM, and it reverts to the stochastic gradient method when no variable splitting is considered. Therefore, the AL framework in stochastic ADMM extends the original stochastic gradient method so that it can use variable splitting for more complicated regularizations such as the non-smooth ℓ_1 regularization, but it does not greatly accelerate convergence for problems with smooth regularizers (in which variable splitting is less compelling than in the non-smooth case) like those considered here for low-dose X-ray CT.

In this paper, we focus on solving regularized (weighted) least-squares problems using a linearized AL method (LALM) [21]. We majorize the quadratic AL penalty term, instead of the smooth data-fitting term, in the scaled augmented Lagrangian using a fixed diagonal majorizer, leading to a much simpler OS-accelerable splitting-based algorithm, OS-LALM. For further acceleration, we use a second-order recursive system analysis to design a deterministic downward continuation approach that avoids tedious parameter tuning and provides fast convergence. Experimental results show that the proposed algorithm significantly accelerates the convergence of X-ray CT image reconstruction in early iterations with negligible overhead and greatly reduces OS artifacts in the reconstructed image when using many subsets.

The paper is organized as follows. Section II reviews the linearized AL method in a general setting and shows new convergence properties of the linearized AL method with inexact updates. Section III derives the proposed OS-accelerable splitting-based algorithm for solving regularized least-squares problems using the linearized AL method and develops a deterministic downward continuation approach for fast convergence without parameter tuning. Section IV considers solving X-ray CT image reconstruction problem with penalized weighted least-squares (PWLS) criterion using the proposed algorithm. Section V reports the experimental results of applying our proposed algorithm to X-ray CT image reconstruction. Further results are shown in the supplementary material. Finally, we draw conclusions in Section VI.

II. Background

A. Linearized AL method (LALM)

Consider a general composite convex optimization problem:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \left\{ g(\mathbf{A}\mathbf{x}) + h(\mathbf{x}) \right\} \tag{1}$$

and its equivalent constrained minimization problem:

$$(\hat{\mathbf{x}}, \hat{\mathbf{u}}) \in \arg\min_{\mathbf{x},\mathbf{u}} \left\{ g(\mathbf{u}) + h(\mathbf{x}) \right\} \text{ s.t. } \mathbf{u} = \mathbf{A}\mathbf{x}\,, \tag{2}$$

where both g and h are closed and proper convex functions. In CT, $\mathbf{A}$ denotes the system (projection) matrix, $\mathbf{x}$ denotes the image being reconstructed, g is a weighted quadratic data-fitting term, and h is an edge-preserving regularization term. One way to solve the constrained minimization problem (2) is to use an (alternating direction) AL method that alternatingly minimizes the scaled augmented Lagrangian:

$$\mathcal{L}_{\mathrm{A}}(\mathbf{x}, \mathbf{u}, \mathbf{d}; \rho) \triangleq g(\mathbf{u}) + h(\mathbf{x}) + \tfrac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{u} - \mathbf{d} \right\|_2^2 \tag{3}$$

with respect to $\mathbf{x}$ and $\mathbf{u}$, followed by a gradient ascent of $\mathbf{d}$, yielding the following AL iterates [4, 6]:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)}\,, \end{cases} \tag{4}$$

where $\mathbf{d}$ is the scaled Lagrange multiplier of the split variable $\mathbf{u}$, and $\rho > 0$ is the corresponding AL penalty parameter.

In the linearized AL method (LALM) [21] (also known as the split inexact Uzawa method [22–24]), one replaces the quadratic AL penalty term in the $\mathbf{x}$-update of (4):

$$\theta_k(\mathbf{x}) \triangleq \tfrac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)} \right\|_2^2 \tag{5}$$

by its separable quadratic surrogate (SQS) function:

$$\begin{aligned} &\breve{\theta}_k\big(\mathbf{x}; \mathbf{x}^{(k)}\big) \\ &\triangleq \theta_k\big(\mathbf{x}^{(k)}\big) + \big\langle \nabla\theta_k\big(\mathbf{x}^{(k)}\big), \mathbf{x} - \mathbf{x}^{(k)} \big\rangle + \tfrac{\rho L}{2} \left\| \mathbf{x} - \mathbf{x}^{(k)} \right\|_2^2 \\ &= \tfrac{\rho}{2t} \left\| \mathbf{x} - \big(\mathbf{x}^{(k)} - t\mathbf{A}'\big(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)}\big)\big) \right\|_2^2 \\ &\qquad\qquad + (\text{constant independent of } \mathbf{x})\,, \end{aligned} \tag{6}$$

where $L > \|\mathbf{A}\|_2^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A})$ ensures $L\mathbf{I} - \mathbf{A}'\mathbf{A} \succ 0$, and $t \triangleq 1/L$. This function satisfies the "majorization" condition:

$$\begin{cases} \breve{\theta}_k(\mathbf{x}; \bar{\mathbf{x}}) \geq \theta_k(\mathbf{x}), & \forall \mathbf{x}, \bar{\mathbf{x}} \in \operatorname{Dom}\theta_k \\ \breve{\theta}_k(\bar{\mathbf{x}}; \bar{\mathbf{x}}) = \theta_k(\bar{\mathbf{x}}), & \forall \bar{\mathbf{x}} \in \operatorname{Dom}\theta_k\,. \end{cases} \tag{7}$$

It is trivial to generalize L to a symmetric positive semi-definite matrix $\mathbf{S}_+$, e.g., the diagonal matrix $\mathbf{D}_{\mathrm{L}}$ used in OS-based algorithms [8, 25], and still ensure (7). When $\mathbf{S}_+ = \mathbf{A}'\mathbf{A}$, LALM just reverts to the standard AL method. Majorizing with a diagonal matrix leads to a simpler $\mathbf{x}$-update. The corresponding LALM iterates are as follows [21]:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ \phi_k(\mathbf{x}) \triangleq h(\mathbf{x}) + \breve{\theta}_k\big(\mathbf{x}; \mathbf{x}^{(k)}\big) \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)}\,. \end{cases} \tag{8}$$

The $\mathbf{x}$-update can be written as the proximal mapping of h:

$$\begin{aligned} \mathbf{x}^{(k+1)} &\in \operatorname{prox}_{(\rho^{-1}t)h}\big(\mathbf{x}^{(k)} - t\mathbf{A}'\big(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)}\big)\big) \\ &= \operatorname{prox}_{(\rho^{-1}t)h}\big(\mathbf{x}^{(k)} - (\rho^{-1}t)\,\mathbf{s}^{(k+1)}\big)\,, \end{aligned} \tag{9}$$

where prox_f denotes the proximal mapping of f defined as:

$$\operatorname{prox}_f(\mathbf{z}) \triangleq \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \tfrac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 \right\} , \tag{10}$$

and

$$\mathbf{s}^{(k+1)} \triangleq \rho \mathbf{A}' \big(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)} \big) \tag{11}$$

denotes the "search direction" of the proximal gradient-like $\mathbf{x}$-update that consists of a descent step and a proximal mapping step using the same step size, for instance, $\rho^{-1}t$ in (9). Furthermore, $\breve{\theta}_k$ can be written as:

$$\breve{\theta}_k \big(\mathbf{x}; \mathbf{x}^{(k)} \big) = \theta_k(\mathbf{x}) + \tfrac{\rho}{2} \left\| \mathbf{x} - \mathbf{x}^{(k)} \right\|_{\mathbf{G}}^2 , \tag{12}$$

where $\mathbf{G} \triangleq L\mathbf{I} - \mathbf{A}'\mathbf{A} \succ 0$ by the definition of L. Hence, the LALM iterates (8) can be represented as a proximal-point variant [24] of the standard AL iterates (4) (also known as the preconditioned ADMM iterates [26] discussed later) by plugging (12) into (8):

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \theta_k(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{x} - \mathbf{x}^{(k)} \right\|_{\mathbf{G}}^2 \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} . \end{cases} \tag{13}$$

B. Convergence properties with inexact updates

The LALM iteration (8) is convergent for any fixed AL penalty parameter $\rho > 0$ and any $\mathbf{A}$ [21], while the standard AL method is convergent (in the primal) if $\mathbf{A}$ has full column rank [4, Theorem 8]. Furthermore, even if the AL penalty parameter varies every iteration, (8) is convergent when ρ is non-decreasing and bounded above [21]. However, existing convergence analyses of LALM assume that all updates are exact. In this paper, since some updates might not be solved exactly, we must consider the LALM iteration with inexact updates. Specifically, instead of the exact LALM in (8), we focus on two closely related inexact LALM variants:

$$\begin{cases} \left\| \mathbf{x}^{(k+1)} - \arg\min_{\mathbf{x}} \phi_k(\mathbf{x}) \right\| \le \delta_k \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} , \end{cases} \tag{14}$$

where ϕ_k was defined in (8), and

$$\begin{cases} \left| \phi_k \big(\mathbf{x}^{(k+1)} \big) - \min_{\mathbf{x}} \phi_k(\mathbf{x}) \right| \le \varepsilon_k \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} . \end{cases} \tag{15}$$

These inexact variants of LALM revert to the standard LALM when $\delta_k = 0$ and $\varepsilon_k = 0$. The $\mathbf{u}$-update could also be inexact; however, for simplicity, we focus on exact updates of $\mathbf{u}$. Considering inexact updates of $\mathbf{u}$ is a trivial extension.

Our convergence analysis of the inexact LALM is twofold. First, we show that the equivalent proximal-point variant of the standard AL iterates (13) can be interpreted as a convergent ADMM that solves another equivalent constrained minimization problem of the form (1) with a redundant split (the proof is in the supplementary material):

$$(\hat{\mathbf{x}}, \hat{\mathbf{u}}, \hat{\mathbf{v}}) \in \arg\min_{\mathbf{x},\mathbf{u},\mathbf{v}} \left\{ g(\mathbf{u}) + h(\mathbf{x}) \right\} \quad \text{s.t. } \mathbf{u} = \mathbf{A}\mathbf{x} \text{ and } \mathbf{v} = \mathbf{G}^{1/2}\mathbf{x} . \tag{16}$$

Therefore, the LALM iteration (8) is a convergent ADMM, and it inherits the nice properties of ADMM, including the tolerance of inexact updates [4, Theorem 8]. More formally, we have the following theorem:

Theorem 1. *Consider a constrained composite convex optimization problem (2) where both g and h are closed and proper convex functions. Let $\rho > 0$ and $\{\delta_k\}_{k=0}^{\infty}$ denote a non-negative sequence such that*

$$\sum_{k=0}^{\infty} \delta_k < \infty . \tag{17}$$

If (2) has a solution $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$, then the sequence of updates $\left\{ \big(\mathbf{x}^{(k)}, \mathbf{u}^{(k)} \big) \right\}_{k=0}^{\infty}$ generated by the inexact LALM in (14) converges to $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$; otherwise, at least one of the sequences $\left\{ \big(\mathbf{x}^{(k)}, \mathbf{u}^{(k)} \big) \right\}_{k=0}^{\infty}$ or $\left\{ \mathbf{d}^{(k)} \right\}_{k=0}^{\infty}$ diverges.

Theorem 1 shows that the inexact LALM in (14) converges if the error δ_k is absolutely summable. However, it does not describe how fast the iterates converge and more importantly, how inexact updates affect the convergence rate. This leads to the second part of our convergence analysis.

In this part, we rely on the equivalence between LALM and the Chambolle-Pock first-order primal-dual algorithm (CP) [26]. Consider a minimax problem:

$$(\hat{\mathbf{z}}, \hat{\mathbf{x}}) \in \arg\min_{\mathbf{z}} \max_{\mathbf{x}} \Omega(\mathbf{z}, \mathbf{x}) , \tag{18}$$

where

$$\Omega(\mathbf{z}, \mathbf{x}) \triangleq \langle -\mathbf{A}'\mathbf{z}, \mathbf{x} \rangle + g^*(\mathbf{z}) - h(\mathbf{x}) , \tag{19}$$

and f^* denotes the convex conjugate of a function f [27, p. 104]. Note that since both g and h are closed, proper, and convex, it follows $g^{**} = g$ and $h^{**} = h$. The sequence of updates $\left\{ \big(\mathbf{z}^{(k)}, \mathbf{x}^{(k)} \big) \right\}_{k=0}^{\infty}$ generated by the CP iterates:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \operatorname{prox}_{\sigma h} \big(\mathbf{x}^{(k)} - \sigma \mathbf{A}' \bar{\mathbf{z}}^{(k)} \big) \\ \mathbf{z}^{(k+1)} \in \operatorname{prox}_{\tau g^*} \big(\mathbf{z}^{(k)} + \tau \mathbf{A}\mathbf{x}^{(k+1)} \big) \\ \bar{\mathbf{z}}^{(k+1)} = \mathbf{z}^{(k+1)} + \big(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)} \big) \end{cases} \tag{20}$$

converges to a saddle-point $(\hat{\mathbf{z}}, \hat{\mathbf{x}})$ of (18), and the non-negative primal-dual gap $\Omega\big(\mathbf{z}_k, \hat{\mathbf{x}}\big) - \Omega\big(\hat{\mathbf{z}}, \mathbf{x}_k\big)$ converges to zero with rate $O(1/k)$ [26, Theorem 1] provided that $\sigma\tau \|\mathbf{A}\|_2^2 < 1$, where $\mathbf{z}_k \triangleq \frac{1}{k} \sum_{j=1}^{k} \mathbf{z}^{(j)}$ and $\mathbf{x}_k \triangleq \frac{1}{k} \sum_{j=1}^{k} \mathbf{x}^{(j)}$ denote the arithemetic means of all previous $\mathbf{z}$- and $\mathbf{x}$-iterates. Since the CP iterates (20) solve the minimax problem (18), they also solve the primal problem:

$$\hat{\mathbf{z}} \in \arg\min_{\mathbf{z}} \left\{ h^*(-\mathbf{A}'\mathbf{z}) + g^*(\mathbf{z}) \right\} \tag{21}$$

and the dual problem:

$$\hat{\mathbf{x}} \in \arg\max_{\mathbf{x}} \left\{ -g(\mathbf{A}\mathbf{x}) - h(\mathbf{x}) \right\} \tag{22}$$

of (18), namely the composite convex optimization problem (1). Therefore, the CP iterates (20) solve (1) with rate

$O(1/k)$ in an ergodic sense, i.e., with respect to $\mathbf{z}_k$ and $\mathbf{x}_k$ instead of $\mathbf{z}^{(k)}$ and $\mathbf{x}^{(k)}$. Furthermore, Chambolle *et al.* showed that their primal-dual algorithm is equivalent to a preconditioned ADMM. For example, the CP iteration (20) for solving (1) is a proximal-point variant of ADMM with a proximal term weighted by $\mathbf{M} \triangleq \sigma^{-1}\mathbf{I} - \tau\mathbf{A}'\mathbf{A}$ provided that $0 < \sigma\tau \left\|\mathbf{A}\right\|_2^2 < 1$ [26, Section 4.3]. Letting $\mathbf{z}^{(k)} = -\tau\mathbf{d}^{(k)}$ and $\bar{\mathbf{z}}^{(0)} = \mathbf{z}^{(0)}$ and choosing $\sigma = \rho^{-1}t$ and $\tau = \rho$, the CP iteration (20) is just the proixmal-point AL method (13) and hence LALM (8) if we initialize $\mathbf{u}$ as $\mathbf{u}^{(0)} = \mathbf{A}\mathbf{x}^{(0)}$. This suggests that we can measure the convergence rate of LALM using the primal-dual gap that vanishes ergodically with rate $O(1/k)$. Finally, to consider inexact updates, we apply the error analysis technique developed in [17] to the convergence rate analysis of CP, leading to the following theorem (the proof is in the supplementary material):

Theorem 2. *Consider a minimax problem (18) where both g and h are closed and proper convex functions. Suppose it has a saddle-point $(\hat{\mathbf{z}}, \hat{\mathbf{x}})$, where $\hat{\mathbf{z}}$ and $\hat{\mathbf{x}}$ are the solutions of the primal problem (21) and the dual problem (22) of (18), respectively. Let $\rho > 0$ and $\{\varepsilon_k\}_{k=0}^{\infty}$ denote a non-negative sequence such that*

$$\sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty \,. \tag{23}$$

Then, the sequence of updates $\left\{\left(-\rho\mathbf{d}^{(k)}, \mathbf{x}^{(k)}\right)\right\}_{k=0}^{\infty}$ generated by the inexact LALM in (15) is a bounded sequence that converges to $(\hat{\mathbf{z}}, \hat{\mathbf{x}})$, and the primal-dual gap of $(\mathbf{z}_k, \mathbf{x}_k)$ has the following bound:

$$\Omega\left(\mathbf{z}_k, \hat{\mathbf{x}}\right) - \Omega\left(\hat{\mathbf{z}}, \mathbf{x}_k\right) \le \frac{\left(C + 2A_k + \sqrt{B_k}\right)^2}{k} \,, \tag{24}$$

where $\mathbf{z}_k \triangleq \frac{1}{k}\sum_{j=1}^{k}\left(-\rho\mathbf{d}^{(j)}\right)$, $\mathbf{x}_k \triangleq \frac{1}{k}\sum_{j=1}^{k}\mathbf{x}^{(j)}$,

$$C \triangleq \frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\rho^{-1}t}} + \frac{\left\|\left(-\rho\mathbf{d}^{(0)}\right) - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\rho}} \,, \tag{25}$$

$$A_k \triangleq \sum_{j=1}^{k} \sqrt{\frac{\varepsilon_{j-1}}{\left(1 - t\left\|\mathbf{A}\right\|_2^2\right)\rho^{-1}t}} \,, \tag{26}$$

and

$$B_k \triangleq \sum_{j=1}^{k} \varepsilon_{j-1} \,. \tag{27}$$

Theorem 2 shows that the inexact LALM in (15) converges with rate $O(1/k)$ if the square root of the error ε_k is absolutely summable. In fact, even if $\left\{\sqrt{\varepsilon_k}\right\}_{k=0}^{\infty}$ is not absolutely summable, say, $\sqrt{\varepsilon_k}$ decreases as $O(1/k)$, A_k grows as $O(\log k)$ (note that B_k always grows slower than A_k), and the primal-dual gap converges to zero in $O\left(\log^2 k/k\right)$. To obtain convergence of the primal-dual gap, a necessary condition is that the partial sum of $\left\{\sqrt{\varepsilon_k}\right\}_{k=0}^{\infty}$ grows no faster than $o\left(\sqrt{k}\right)$.

The primal-dual gap convergence bound above is measured at the average point $(-\rho\mathbf{d}_k, \mathbf{x}_k)$ of the update trajectory. In practice, the primal-dual gap of $\left(-\rho\mathbf{d}^{(k)}, \mathbf{x}^{(k)}\right)$ converges much faster. Minimizing the constant in (24) need not provide the fastest convergence rate. However, the ρ-, t-, and ε_k-dependence in (24) suggests how these factors affect the convergence rate of LALM. Note that although we consider only one variable split in our derivation, it is easy to extend our proofs to support multiple variable splits by using the variable splitting scheme in [6]. Conventional LALM in (8) is not OS-accelerable because it needs one full forward projection for the $\mathbf{u}$- and $\mathbf{d}$-updates. We used LALM for analysis and to motivate the proposed algorithm in Section III, but it is not recommended for practical implementation in CT reconstruction. By restricting g to be a quadratic loss function, we show next that LALM becomes OS-accelerable and can further accelerate the conventional OS algorithms by decreasing or choosing a small AL penalty paremeter.

III. Proposed algorithm

A. OS-LALM: an OS-accelerable splitting-based algorithm

In this section, we restrict g to be a quadratic loss function, i.e., $g(\mathbf{u}) \triangleq \frac{1}{2}\left\|\mathbf{y} - \mathbf{u}\right\|_2^2$, and then the minimization problem (1) becomes a regularized least-squares problem:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \left\{\Psi(\mathbf{x}) \triangleq \tfrac{1}{2}\left\|\mathbf{y} - \mathbf{A}\mathbf{x}\right\|_2^2 + h(\mathbf{x})\right\} . \tag{28}$$

Let $\mathsf{L}(\mathbf{x}) \triangleq g(\mathbf{A}\mathbf{x})$ denote the quadratic data-fitting term in (28). We assume that L is suitable for OS acceleration; i.e., L can be decomposed into M smaller quadratic functions $\mathsf{L}_1, \ldots, \mathsf{L}_M$ satisfying the "subset balance condition" [8]:

$$\nabla\mathsf{L}(\mathbf{x}) \approx M\nabla\mathsf{L}_1(\mathbf{x}) \approx \cdots \approx M\nabla\mathsf{L}_M(\mathbf{x}) \ , \tag{29}$$

so that the subset gradients approximate the gradient of L.

Since g is quadratic, its proximal mapping is linear. The $\mathbf{u}$-update in LALM (8) has the following simple closed-form solution:

$$\mathbf{u}^{(k+1)} = \tfrac{\rho}{\rho+1}\left(\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{d}^{(k)}\right) + \tfrac{1}{\rho+1}\mathbf{y} \,. \tag{30}$$

Combining (30) with the $\mathbf{d}$-update of (8) yields the identity

$$\mathbf{u}^{(k+1)} + \rho\mathbf{d}^{(k+1)} = \mathbf{y} \tag{31}$$

if we initialize $\mathbf{d}$ as $\mathbf{d}^{(0)} = \rho^{-1}\left(\mathbf{y} - \mathbf{u}^{(0)}\right)$. Letting $\tilde{\mathbf{u}} \triangleq \mathbf{u} - \mathbf{y}$ denote the split residual and substituting (31) into (8) lead to the following simplified LALM iterates:

$$\begin{cases} \mathbf{s}^{(k+1)} = \mathbf{A}'\left(\rho\left(\mathbf{A}\mathbf{x}^{(k)} - \mathbf{y}\right) + (1-\rho)\,\tilde{\mathbf{u}}^{(k)}\right) \\ \mathbf{x}^{(k+1)} \in \operatorname{prox}_{(\rho^{-1}t)h}\left(\mathbf{x}^{(k)} - (\rho^{-1}t)\,\mathbf{s}^{(k+1)}\right) \\ \tilde{\mathbf{u}}^{(k+1)} = \tfrac{\rho}{\rho+1}\left(\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{y}\right) + \tfrac{1}{\rho+1}\tilde{\mathbf{u}}^{(k)} \,. \end{cases} \tag{32}$$

The net computational complexity of (32) per iteration reduces to one multiplication by $\mathbf{A}$, one multiplication by $\mathbf{A}'$, and one proximal mapping of h that often can be solved non-iteratively or solved iteratively without using $\mathbf{A}$ or $\mathbf{A}'$. Since the gradient of L is $\mathbf{A}'\left(\mathbf{A}\mathbf{x} - \mathbf{y}\right)$, letting $\mathbf{g} \triangleq \mathbf{A}'\tilde{\mathbf{u}}$ (a back-projection of the split residual) denote the split gradient, we rewrite (32) as:

$$\boxed{\begin{cases} \mathbf{s}^{(k+1)} = \rho\nabla\mathsf{L}\left(\mathbf{x}^{(k)}\right) + (1-\rho)\,\mathbf{g}^{(k)} \\ \mathbf{x}^{(k+1)} \in \operatorname{prox}_{(\rho^{-1}t)h}\left(\mathbf{x}^{(k)} - (\rho^{-1}t)\,\mathbf{s}^{(k+1)}\right) \\ \mathbf{g}^{(k+1)} = \tfrac{\rho}{\rho+1}\nabla\mathsf{L}\left(\mathbf{x}^{(k+1)}\right) + \tfrac{1}{\rho+1}\mathbf{g}^{(k)} \,. \end{cases}} \tag{33}$$

We call (33) the gradient-based LALM because only the gradients of L are used to perform the updates, and the net computational complexity of (33) per iteration becomes one gradient evaluation of L and one proximal mapping of h.

We interpret the gradient-based LALM (33) as a generalized proximal gradient descent of a regularized least-squares cost function Ψ with step size $\rho^{-1}t$ and search direction $\mathbf{s}^{(k+1)}$ that is a weighted average of the gradient and split gradient of L. A smaller ρ can lead to a larger step size. When $\rho = 1$, (33) happens to be the proximal gradient method or the iterative shrinkage/thresholding algorithm (ISTA) [28]. In other words, by using LALM, we can arbitrarily increase the step size of the proximal gradient method by decreasing ρ, thanks to the simple ρ-dependent correction of the search direction in (33). To have a concrete example, suppose all updates are exact, i.e., $\varepsilon_k = 0$ for all k. From (31) and Theorem 2, we have $-\rho\mathbf{d}^{(k)} = \mathbf{u}^{(k)} - \mathbf{y} \to \mathbf{A}\hat{\mathbf{x}} - \mathbf{y} = \hat{\mathbf{z}}$ as $k \to \infty$. Furthermore, $\left(-\rho\mathbf{d}^{(0)}\right) - \hat{\mathbf{z}} = \mathbf{u}^{(0)} - \mathbf{A}\hat{\mathbf{x}}$. With a reasonable initialization, e.g., $\mathbf{u}^{(0)} = \mathbf{A}\mathbf{x}^{(0)}$ and consequently, $\mathbf{g}^{(0)} = \nabla\mathsf{L}\left(\mathbf{x}^{(0)}\right)$, the constant C in (25) can be rewritten as a function of ρ:

$$C(\rho) = \frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\rho^{-1}t}} + \frac{\left\|\mathbf{A}\left(\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right)\right\|_2}{\sqrt{2\rho}}\,. \tag{34}$$

This constant achieves its minimum at

$$\rho_{\text{opt}} = \frac{\left\|\mathbf{A}\left(\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right)\right\|_2}{\sqrt{L}\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2} \leq 1\,, \tag{35}$$

suggesting that unity might be a reasonable upper bound on ρ for fast convergence. Note that the ratio of the first term to the second term in (34) is ρ/ρ_{opt}. When the majorization is loose, i.e., $L \gg \|\mathbf{A}\|_2^2$, $\rho_{\text{opt}} \ll 1$ and the first term in (34) dominates C for $\rho_{\text{opt}} < \rho \leq 1$ since $\rho/\rho_{\text{opt}} \gg 1$. The upper bound of the primal-dual gap becomes

$$\Omega\left(\mathbf{z}_k, \hat{\mathbf{x}}\right) - \Omega\left(\hat{\mathbf{z}}, \mathbf{x}_k\right) \leq \frac{C^2}{k} \approx \frac{\frac{L}{2}\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{\rho^{-1}k}\,. \tag{36}$$

That is, comparing to the proximal gradient method ($\rho = 1$), the convergence rate (bound) of (33) is accelerated by a factor of ρ^{-1} for $\rho_{\text{opt}} < \rho \leq 1$.

Finally, since (33) requires only the gradients of L to perform updates, it is OS-accelerable. For OS acceleration, we simply replace $\nabla\mathsf{L}$ in (33) with $M\nabla\mathsf{L}_m$ using the approximation (29) and incrementally perform (33) for M times as a complete iteration, thus leading to the final proposed OS-accelerable LALM (OS-LALM):

$$\boxed{\begin{cases} \mathbf{s}^{(k,m+1)} = \rho M\nabla\mathsf{L}_m\left(\mathbf{x}^{(k,m)}\right) + (1-\rho)\,\mathbf{g}^{(k,m)} \\ \mathbf{x}^{(k,m+1)} \in \operatorname{prox}_{(\rho^{-1}t)h}\left(\mathbf{x}^{(k,m)} - \left(\rho^{-1}t\right)\mathbf{s}^{(k,m+1)}\right) \\ \mathbf{g}^{(k,m+1)} = \frac{\rho}{\rho+1}M\nabla\mathsf{L}_{m+1}\left(\mathbf{x}^{(k,m+1)}\right) + \frac{1}{\rho+1}\mathbf{g}^{(k,m)} \end{cases}} \tag{37}$$

with $\mathbf{c}^{(k,M+1)} = \mathbf{c}^{(k+1)} = \mathbf{c}^{(k+1,1)}$ for $\mathbf{c} \in \{\mathbf{s}, \mathbf{x}, \mathbf{g}\}$ and $\mathsf{L}_{M+1} = \mathsf{L}_1$. Like typical OS-based algorithms, this algorithm is convergent when $M = 1$, i.e., (33), but is not guaranteed to converge for $M > 1$. When $M > 1$, updates generated by OS-based algorithms approach a "limit cycle" in which updates stop nearing the optimum, and visible OS artifacts might be observed in the reconstructed image, depending on M.

B. Deterministic downward continuation

One drawback of conventional LALM is the difficulty of finding a fixed value for the penalty parameter ρ that provides the fastest convergence. The optimal penalty parameter ρ_{opt} in (35) minimizes the multiplicative constant but depends on the unknown solution $\hat{\mathbf{x}}$ of the problem. Intuitively, a smaller ρ is better because it leads to a larger step size. However, when the step size is too large, one can encounter overshoots and oscillations that slow down the convergence rate at first and when nearing the optimum. In fact, ρ_{opt} in (35) also suggests that ρ should not be arbitrarily small. Rather than estimating ρ_{opt} heuristically, we focus on using an iteration-dependent ρ, i.e., a continuation approach, for acceleration.

Classic continuation approaches increase ρ as iterations progress so that previous iterates can serve as a warm start for subsequent worse-conditioned but more penalized inner minimization problems [29, Proposition 4.2.1]. To implement this kind of approaches, one must specify an initial value of the penalty parameter ρ_0 and a rule for increasing ρ which are usually problem-dependent. For example, a small ρ_0 provides fast initial convergence but can cause overshoot (e.g., increasing the cost function) in early iterations. Choosing a good ρ_0 that balances convergence rate and overshoot for a given problem is difficult. Furthermore, most classic continuation approaches adapt the penalty parameter dynamically by checking some conditions such as the primal and dual feasibilities and the decrease of cost function [30]. This can increase computational complexity per iteration, especially when involving expensive operations (e.g., $\mathbf{A}$ and $\mathbf{A}'$ in our case).

In this paper, unlike classic continuation approaches, we consider a downward continuation approach. It is inspired by Nesterov's second method [31] that starts as a proximal gradient method and gradually increases the step size (of the auxiliary sequence) deterministically. The intuition is that, for a fixed ρ, the step length $\left\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\right\|$ is typically a decreasing sequence because the gradient norm vanishes as we approach the optimum, and an increasing sequence ρ_k (i.e., a diminishing step size) would aggravate the shrinkage of step length, slowing convergence. In contrast, a decreasing sequence ρ_k can compensate for step length shrinkage and accelerate convergence. Of course, ρ_k cannot decrease too fast; otherwise, the soaring step size might make the algorithm unstable or even divergent. To design a "good" decreasing sequence ρ_k for "effective" acceleration, we first analyze how LALM (the one-subset version (33) for simplicity) behaves for different values of ρ.

Consider a simple quadratic problem:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \tfrac{1}{2}\left\|\mathbf{A}\mathbf{x}\right\|_2^2\,, \tag{38}$$

corresponding to (28) with $h = 0$ and $\mathbf{y} = \mathbf{0}$. A trivial solution of (38) is $\hat{\mathbf{x}} = \mathbf{0}$. To ensure a unique solution, we assume that $\mathbf{A}'\mathbf{A}$ is positive definite (for this analysis only). Let $\mathbf{A}'\mathbf{A}$ have eigenvalue decomposition $\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'$, where $\boldsymbol{\Lambda} \triangleq \mathsf{diag}\{\lambda_i\}$ and $0 < \lambda_1 \leq \cdots \leq \lambda_n = L$. The updates generated by (33) that solve (38) can be written as

$$\begin{cases}\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - (1/L)\big(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'\mathbf{x}^{(k)} + (\rho^{-1}-1)\,\mathbf{g}^{(k)}\big)\\ \mathbf{g}^{(k+1)} = \frac{\rho}{\rho+1}\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}'\mathbf{x}^{(k+1)} + \frac{1}{\rho+1}\mathbf{g}^{(k)}\,.\end{cases} \tag{39}$$

Furthermore, letting $\bar{\mathbf{x}} = \mathbf{V}'\mathbf{x}$ and $\bar{\mathbf{g}} = \mathbf{V}'\mathbf{g}$, the linear system can be further diagonalized, and we can represent the ith components of $\bar{\mathbf{x}}$ and $\bar{\mathbf{g}}$ as

$$\begin{cases}\bar{x}_i^{(k+1)} = \bar{x}_i^{(k)} - (1/L)\big(\lambda_i\bar{x}_i^{(k)} + (\rho^{-1}-1)\,\bar{g}_i^{(k)}\big)\\ \bar{g}_i^{(k+1)} = \frac{\rho}{\rho+1}\lambda_i\bar{x}_i^{(k+1)} + \frac{1}{\rho+1}\bar{g}_i^{(k)}\,.\end{cases} \tag{40}$$

Solving this system of recurrence relations of $\bar{x}_i$ and $\bar{g}_i$, one can show that both $\bar{x}_i$ and $\bar{g}_i$ satisfy a second-order recursive system determined by the characteristic polynomial:

$$(1+\rho)\,r_i^2 - 2\,(1-\lambda_i/L+\rho/2)\,r_i + (1-\lambda_i/L)\,. \tag{41}$$

The roots r_i of this polynomial determine the convergence rate of $\bar{x}_i$ and $\bar{g}_i$ in (40).

When $\rho = \rho_i^{\mathrm{c}}$, where

$$\rho_i^{\mathrm{c}} \triangleq 2\sqrt{\frac{\lambda_i}{L}\left(1-\frac{\lambda_i}{L}\right)} \in (0,1]\,, \tag{42}$$

the characteristic polynomial (41) has repeated roots. Hence, the system is critically damped, and $\bar{x}_i$ and $\bar{g}_i$ converge geometrically to zero with convergence rate

$$r_i^{\mathrm{c}} = \frac{1-\lambda_i/L+\rho_i^{\mathrm{c}}/2}{1+\rho_i^{\mathrm{c}}} = \sqrt{\frac{1-\lambda_i/L}{1+\rho_i^{\mathrm{c}}}}\,. \tag{43}$$

When $\rho > \rho_i^{\mathrm{c}}$, the characteristic polynomial (41) has distinct real roots. Hence, the system is over-damped, and $\bar{x}_i$ and $\bar{g}_i$ converge geometrically to zero with convergence rate that is governed by the dominant root

$$r_i^{\mathrm{o}}(\rho) = \frac{1-\lambda_i/L+\rho/2+\sqrt{\rho^2/4-\lambda_i/L\,(1-\lambda_i/L)}}{1+\rho}\,. \tag{44}$$

It is easy to check that $r_i^{\mathrm{o}}(\rho_i^{\mathrm{c}}) = r_i^{\mathrm{c}}$, and r_i^{o} is increasing. This suggests that the critically damped system always converges faster than the over-damped system. Finally, when $\rho < \rho_i^{\mathrm{c}}$, the characteristic polynomial (41) has complex roots. In this case, the system is under-damped, and $\bar{x}_i$ and $\bar{g}_i$ converge geometrically to zero with convergence rate

$$r_i^{\mathrm{u}}(\rho) = \frac{1-\lambda_i/L+\rho/2}{1+\rho}\,, \tag{45}$$

and oscillate at the damped frequency $\psi_i/(2\pi)$, where

$$\cos\psi_i = \frac{1-\lambda_i/L+\rho/2}{\sqrt{(1+\rho)(1-\lambda_i/L)}} \approx \sqrt{1-\lambda_i/L} \tag{46}$$

when $\rho \approx 0$. Furthermore, by the small angle approximation: $\cos\sqrt{\theta} \approx 1-\theta/2 \approx \sqrt{1-\theta}$, if $\lambda_i \ll L$, $\psi_i \approx \sqrt{\lambda_i/L}$. Again, $r_i^{\mathrm{u}}(\rho_i^{\mathrm{c}}) = r_i^{\mathrm{c}}$, but r_i^{u} behaves differently from r_i^{o}. Specifically, r_i^{u} is decreasing if $\lambda_i/L < 1/2$, and it is increasing otherwise. This suggests that the critically damped system converges faster than the under-damped system if $\lambda_i/L < 1/2$, but it can be slower otherwise. In sum, the critically damped system is optimal for eigencomponents having smaller eigenvalues (i.e., $\lambda_i < L/2$), while for eigencomponents having larger eigenvalues (i.e., $\lambda_i > L/2$), the under-damped system is optimal.

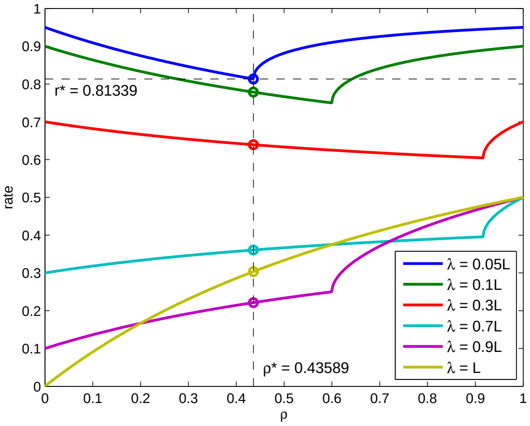

Fig. 1: The dominant roots $r_i(\rho)$ of (41) and the optimal asymptotic convergence rate $r^\star \triangleq \min_\rho\{\max_i r_i(\rho)\}$ for a system with six distinct eigenvalues: $0.05L$, $0.1L$, $0.3L$, $0.7L$, $0.9L$, and L.

The asymptotic convergence rate of the system is dominated by the smallest eigenvalue λ_1. Eventually, only the component oscillating at frequency $\psi_1/(2\pi)$ persists. Therefore, for the fastest asymptotic convergence rate, we would like to choose the AL penalty parameter ρ to be

$$\rho^\star = \rho_1^{\mathrm{c}} = 2\sqrt{\frac{\lambda_1}{L}\left(1-\frac{\lambda_1}{L}\right)} \in (0,1]\,. \tag{47}$$

Figure 1 illustrates the dominant roots of (41) and the optimal asymptotic convergence rate that is the minimum of the largest dominant root, i.e., $\max_i r_i(\rho)$, of all possible ρ for a system with six distinct eigenvalues: $0.05L$, $0.1L$, $0.3L$, $0.7L$, $0.9L$, and L. For eigencomponents having smaller eigenvalues ($0.05L$, $0.1L$, and $0.3L$), the critically damped system has the fastest asymptotic convergence rate, while eigencomponents having larger eigenvalues ($0.7L$, $0.9L$, and L) attain the fastest asymptotic convergence rate in the under-damping regime. Moreover, when the smallest eigenvalue is less than $L/2$ (i.e., an ill-conditioned system), the eigencomponent with the smallest eigenvalue determines the optimal asymptotic convergence rate, i.e., $\min_\rho\{\max_i r_i(\rho)\}$, with $\rho^\star$ in (47). Unlike ρ_{opt} in (35), this choice of ρ does not depend on the initialization. It depends only on the geometry of the Hessian $\mathbf{A}'\mathbf{A}$. Furthermore, both ρ_{opt} and $\rho^\star$ fall in the interval $(0,1]$. Hence, although LALM converges for any $\rho > 0$, we consider only $0 < \rho \le 1$ in our downward continuation approach.

We can now interpret classic (upward) continuation approaches based on the second-order recursive system analysis. Classic continuation approaches usually start from a small ρ for better-conditioned inner minimization problem. Therefore, initially, the system is under-damped. Although the under-damped system has a slower asymptotic convergence rate, the oscillation can provide dramatic acceleration before the first zero-crossing of the oscillating components. We can think of the classic continuation approach as a greedy strategy that exploits the initial fast convergence rate of the under-damped system and carefully increases ρ to avoid oscillation and move toward the critical damping regime. However, this greedy strategy requires a "clever" update rule for increasing ρ. If ρ increases too fast, the acceleration ends prematurally; if ρ increases too slow, the system starts oscillating.

In contrast, we consider a more conservative strategy that starts from the over-damped regime, say, $\rho = 1$ as suggested in (47), and gradually reduces ρ to the optimal AL penalty parameter $\rho^\star$. It sounds impractical at first because we do not know λ_1 beforehand. To solve this problem, we adopt the adaptive restart proposed in [32] and generate a decreasing sequence ρ_k that starts from $\rho = 1$ and reaches $\rho^\star$ every time the algorithm restarts. As mentioned before, the system oscillates at frequency $\psi_1/(2\pi)$ when it is under-damped. This oscillating behavior can also be observed from the trajectory of updates. For example,

$$\xi(k) \triangleq \big(\mathbf{g}^{(k)} - \nabla \mathsf{L}\big(\mathbf{x}^{(k+1)}\big)\big)'\big(\nabla \mathsf{L}\big(\mathbf{x}^{(k+1)}\big) - \nabla \mathsf{L}\big(\mathbf{x}^{(k)}\big)\big) \tag{48}$$

oscillates at the frequency ψ_1/π [32]. Hence, we restart the algorithm (i.e., reset the decreasing penalty parameter ρ to be one and $\mathbf{g}$ to be the current gradient of L) every time $\xi(k) > 0$, which should occur about every $(\pi/2)\sqrt{L/\lambda_1}$ iterations. Suppose we restart at the rth iteration, we have the approximation $\sqrt{\lambda_1/L} \approx \pi/(2r)$, and the ideal AL penalty parameter at the rth iteration should be

$$2\sqrt{\left(\tfrac{\pi}{2r}\right)^2\left(1-\left(\tfrac{\pi}{2r}\right)^2\right)} = \tfrac{\pi}{r}\sqrt{1-\left(\tfrac{\pi}{2r}\right)^2}\,. \tag{49}$$

The proposed downward continuation approach has the form (33), where we replace every ρ in (33) with

$$\rho_l = \begin{cases} 1, & \text{if } l = 0 \\ \max\left\{\frac{\pi}{l+1}\sqrt{1-\left(\frac{\pi}{2l+2}\right)^2}, \rho_{\text{min}}\right\}, & \text{otherwise}\,, \end{cases} \tag{50}$$

where l is a counter that starts from zero, increases by one everytime $\mathbf{g}$ is updated, and is reset to zero whenever $\xi(k) > 0$. The lower bound ρ_{min} is a small positive number for guaranteeing convergence. Note that ADMM is convergent if ρ is non-increasing and bounded below away from zero [33, Corollary 4.2]. As shown in Section II-B (Theorem 1), LALM is a convergent ADMM. Therefore, we can ensure convergence (of the one-subset version) of the proposed downward continuation approach if we set a non-zero lower bound for ρ_l, e.g., $\rho_{\text{min}} = 10^{-3}$ in our experiments. Note that ρ_l in (50) is the same for any $\mathbf{A}$. The adaptive restart condition takes care of the dependence on $\mathbf{A}$. That is why we call this approach the deterministic downward continuation approach. When h is non-zero and/or $\mathbf{A}'\mathbf{A}$ is not positive definite, our analysis above does not hold. However, the deterministic downward continuation approach works well in practice for CT. One possible explanation is that the cost function can usually be well approximated by a quadratic near the optimum when the minimization problem is well-posed and h is locally quadratic.

Finally, in practice we do not restart our algorithm for X-ray CT image reconstruction (with ordered subsets, i.e., (37)) for two reasons. First, according to our analysis, the restart period is proportional to the square root of the local condition number, i.e., $\sqrt{L/\lambda_1}$. Since X-ray CT image reconstruction problems are usually very ill-conditioned, the algorithm usually terminates before a restart is needed. Second, since OS is used for acceleration, gradients used to compute $\xi^{(k)}$ are not accurate and might lead to premature restart. In our experimental results, we did not observe any problems without restart. However, restart may be useful in other applications.

IV. Implementation details

In this section, we consider solving the X-ray CT image reconstruction problem:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\Omega}\left\{\tfrac{1}{2}\left\|\mathbf{y}-\mathbf{A}\mathbf{x}\right\|_{\mathbf{W}}^2 + \mathsf{R}(\mathbf{x})\right\} \tag{51}$$

using the proposed OS-LALM algorithm (37), where $\mathbf{A}$ is the system matrix of a CT scan, $\mathbf{y}$ is the noisy sinogram, $\mathbf{W}$ is the statistical weighting matrix, R is an edge-preserving regularizer, and Ω denotes the convex set for a box constraint (usually the non-negativity constraint) on $\mathbf{x}$. We focus on the edge-preserving regularizer R defined as:

$$\mathsf{R}(\mathbf{x}) \triangleq \sum_i \beta_i \sum_n \kappa_n \kappa_{n+s_i} \phi_i([\mathbf{C}_i\mathbf{x}]_n)\;, \tag{52}$$

where β_i, s_i, ϕ_i, and $\mathbf{C}_i$ denote the regularization parameter, corresponding offset, potential function, and finite difference matrix in the ith direction, respectively, and κ_n is a voxel-dependent weight for improving resolution uniformity [34]. In our experiments, we use 13 directions to include all neighbors in 3D CT.

A. OS-LALM for X-ray CT image reconstruction

The X-ray CT image reconstruction problem (51) is a constrained regularized weighted least-squares problem. To solve it using the proposed algorithm (33) and its OS variant (37), we use the following substitution:

$$\begin{cases} \mathbf{A} \leftarrow \mathbf{W}^{1/2}\mathbf{A} \\ \mathbf{y} \leftarrow \mathbf{W}^{1/2}\mathbf{y} \\ h \leftarrow \mathsf{R} + \iota_\Omega\,, \end{cases} \tag{53}$$

where $\iota_{\mathcal{C}}$ denotes the characteristic function of a convex set $\mathcal{C}$. Thus, the inner minimization problem in (33) and its OS variant (37) becomes a constrained denoising problem. In our implementation, we solve this inner constrained denoising problem using n iterations of the fast iterative shrinkage/thresholding algorithm (FISTA) [15] starting from the previous update as a warm start. As discussed in Section II-B, inexact updates can slow down the convergence rate of the proposed algorithm. In general, the more FISTA iterations, the faster convergence rate of the proposed algorithm. However, the overhead of iterative inner updates is non-negligible for large n, especially when the number of subsets is large. Fortunately, in typical X-ray CT image reconstruction problems, the majorization is usually very loose (probably due to the huge dynamic range of the statistical weighting $\mathbf{W}$). Therefore, $t \ll 1$ in most cases, greatly diminishing the regularization force in the constrained denoising problem. In practice, the constrained denoising problem can be solved up to some acceptable tolerance within just one or two iterations. For a fair comparison with the OS-based algorithms [8, 12] used in Section V that majorize the weighted quadratic data-fitting term L using the SQS function with a diagonal Hessian $\mathbf{D}_{\mathsf{L}} \triangleq \mathsf{diag}\{\mathbf{A}'\mathbf{W}\mathbf{A}\mathbf{1}\}$ [8], in our experiments, we also majorize the quadratic penalty in the scaled augmented Lagrangian using the SQS function with Hessian $\mathbf{D}_{\mathsf{L}}$ (in this case, the inner minimization problem in (33) and its OS variant (37) becomes a constrained weighted denoising problem that can also be solved by FISTA) and incrementally update the image using the subset gradients with the bit-reversal order [35] that heuristically minimizes the subset gradient variance as in other OS-based algorithms. See the supplementary material for the outline of the proposed OS-LALM algorithm for solving (51).

The SQS function with Hessian $\mathbf{D}_{\mathsf{L}}$ is a very loose majorizer. For fastest convergence, one might wish to use the tightest majorizer with Hessian $\mathbf{A}'\mathbf{W}\mathbf{A}$. However, this would revert to the standard AL method (4) with expensive $\mathbf{x}$-updates. An alternative is the Barzilai-Borwein (spectral) method [36] that mimics the Hessian $\mathbf{A}'\mathbf{W}\mathbf{A}$ by $\mathbf{H}_k \triangleq \alpha_k \mathbf{D}_{\mathsf{L}}$, where the scaling factor α_k is solved by fitting the secant equation in the (weighted) least-squares sense. Detailed derivation and additional experimental results can be found in the supplementary material.

B. Number of subsets

As mentioned in Section III-A, the number of subsets M can affect the stability of OS-based algorithms. When M is too large, OS algorithms typically become unstable, causing artifacts in the reconstructed image. Therefore, finding an appropriate number of subsets is very important. Since errors of OS-based algorithms come from the gradient approximation using subset gradients, artifacts might be supressed using a better gradient approximation. Intuitively, to have a reasonable gradient approximation, each voxel in a subset should be sampled by a minimum number of views s. For simplicity, we consider the central voxel in the transaxial plane. In axial CT, the views are uniformly distributed in each subset, so we want

$$\tfrac{1}{M_{\text{axial}}} \cdot (\text{number of views}) \geq s_{\text{axial}} \, . \tag{54}$$

This leads to our maximum number of subsets for axial CT:

$$M_{\text{axial}} \leq (\text{number of views}) \cdot \tfrac{1}{s_{\text{axial}}} \, . \tag{55}$$

Helical CT is more complicated. Since the X-ray source moves in the z direction, a central voxel is only covered by $d_{\text{so}}/(p \cdot d_{\text{sd}})$ turns, where p is the pitch, d_{so} denotes the distance from the X-ray source to the isocenter, and d_{sd} denotes the distance from the X-ray source to the detector. Therefore, we want

$$\tfrac{1}{M_{\text{helical}}} \cdot (\text{number of views per turn}) \cdot \tfrac{d_{\text{so}}}{p \cdot d_{\text{sd}}} \geq s_{\text{helical}} \, . \tag{56}$$

This leads to our maximum number of subsets for helical CT:

$$M_{\text{helical}} \leq (\text{number of views per turn}) \cdot \tfrac{d_{\text{so}}}{p \cdot s_{\text{helical}} \cdot d_{\text{sd}}} \, . \tag{57}$$

Note that the maximum number of subsets for helical CT M_{helical} is inversely proportional to the pitch p. We set $s_{\text{axial}} \approx 40$ and $s_{\text{helical}} \approx 24$ for the proposed algorithm in our experiments.

V. Experimental results

This section reports numerical results for 3D X-ray CT image reconstruction from real CT scans with different geometries using various OS-based algorithms, including

- **OS-SQS-$\boldsymbol{M}$**: the standard OS algorithm [8] with M subsets,
- **OS-Nes05-$\boldsymbol{M}$**: the OS+momentum algorithm [12] based on Nesterov's fast gradient method [14] with M subsets,
- **OS-LALM-$\boldsymbol{M}$-$\boldsymbol{\rho}$-$\boldsymbol{n}$**: the proposed algorithm using a fixed AL penalty parameter ρ with M subsets and n FISTA iterations for solving the inner constrained denoising problem, and
- **OS-LALM-$\boldsymbol{M}$-c-$\boldsymbol{n}$**: the proposed algorithm using the deterministic downward continuation approach described in Section III-B with M subsets and n FISTA iterations for solving the inner constrained denoising problem.

OS-SQS is a standard iterative method for tomographic reconstruction, and OS-Nes05 is a state-of-the-art method for fast X-ray CT image reconstruction using Nesterov's momentum technique. Unlike other OS-based algorithms, our proposed algorithm has additional overhead due to the iterative inner updates. However, when $n = 1$, i.e., with a single gradient descent for the constrained denoising problem, all algorithms listed above have the same computational complexity (one forward/back-projection pair and M regularizer gradient evaluations per iteration). When majorizing the regularizer, we use Huber's curvature [37, p. 185] for faster convergence in all algorithms. Therefore, comparing the convergence rate as a function of iteration is fair. We measured the convergence rate using the RMS difference (in the region of interest) between the reconstructed image $\mathbf{x}^{(k)}$ and the almost converged reference reconstruction $\mathbf{x}^\star$ that we generated by running several iterations of the OS+momentum algorithm with a small M, followed by 2000 iterations of a convergent (i.e., one-subset) FISTA with adaptive restart [32]. We used a q-generalized Gaussian [38] potential function ϕ_i in (52), and β_i and κ_n were tuned to emulate GE's Veo method [39]. For reproducible results using an XCAT phantom, see the supplementary material.

A. Shoulder scan

In this experiment, we reconstructed a 512×512×109 image from a shoulder region helical CT scan, where the sinogram has size 888 × 32 × 7146 and pitch 0.5. The maximum number of subsets suggested by (57) is about 40. Figure 2 shows the cropped images from the central transaxial plane of the initial FBP image, the reference reconstruction, and the reconstructed image using the proposed algorithm (OS-LALM-40-c-1) at the 30th iteration (i.e., after 30 forward/back-projection pairs). In Figure 2, the reconstructed image using the proposed algorithm looks almost the same as the reference reconstruction in the display window from 800 to 1200 Hounsfield unit (HU, modified so that air is 0). The reconstructed image using the OS+momentum algorithm (not shown here) also looks quite similar to the reference reconstruction.

Figure 3 shows the difference images, i.e., $\mathbf{x}^{(30)} - \mathbf{x}^{\star}$, for different OS-based algorithms. The standard OS algorithm (with both 20 and 40 subsets) exhibits visible streak artifacts and structured high frequency noise in the difference image. When $M = 20$, the difference images look similar for the OS+momentum algorithm and our proposed algorithm, although that of the OS+momentum algorithm is slightly more structured and non-uniform. When $M = 40$, the difference image for our proposed algorithm remains uniform, whereas some noise-like OS artifacts appear in the OS+momentum algorithm's difference image. The OS artifacts for the OS+momentum algorithm worsen when M increases, e.g., $M = 80$ (not shown). Apparently OS-LALM has better gradient error tolerance than previous OS methods, probably due to the way we compute the search direction and the less aggressive acceleration to the regularization term. Additional experimental results (an XCAT phantom axial scan and a truncated abdomen scan) in the supplementary material demonstrate how different OS-based algorithms behave when M exceeds the suggested maximum number of subsets.

Figure 4 shows the convergence rate curves (RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration) using OS-based algorithms with 20 and 40 subsets. By exploiting the linearized AL method, the proposed algorithm accelerates the standard OS algorithm remarkably. As mentioned in Section III-A, a smaller ρ can provide greater acceleration due to the increased step size. Both plots show the acceleration of convergence as ρ decreases. Note that too large step sizes can cause overshoots in early iterations. For example, the proposed algorithm with $\rho = 0.05$ shows slower convergence rate in first few iterations but decreases more rapidly later. Our proposed deterministic downward continuation approach (50) overcomes this tradeoff. In Figure 4, the proposed algorithm using deterministic downward continuation reaches the lowest RMSD (lower than 1 HU) within only 30 iterations. The slightly higher RMSD of the OS+momentum algorithm with 40 subsets is due to the OS artifacts seen in Figure 3.

Figure 5 illustrates the effectiveness of solving the inner constrained denoising problem using FISTA (for X-ray CT image reconstruction) mentioned in Section IV-A. In Figure 5, the convergence rate improves only slightly when using more than one FISTA iteration for solving the inner constrained denoising problem. In practice, one FISTA iteration, i.e., $n = 1$, per subset update suffices for fast and accurate X-ray CT image reconstruction.

B. GE performance phantom

In this experiment, we reconstructed a 1024 × 1024 × 90 image from the GE performance phantom (GEPP) axial CT scan, where the sinogram has size 888 × 64 × 984. The maximum number of subsets suggested by (55) is about 24. Figure 6 shows the cropped images from the central transaxial plane of the initial FBP image, the reference reconstruction, and the reconstructed image using the proposed algorithm (OS-LALM-24-c-1) at the 30th iteration. Again, the OS-LALM image at the 30th iteration is very similar to the reference reconstruction.

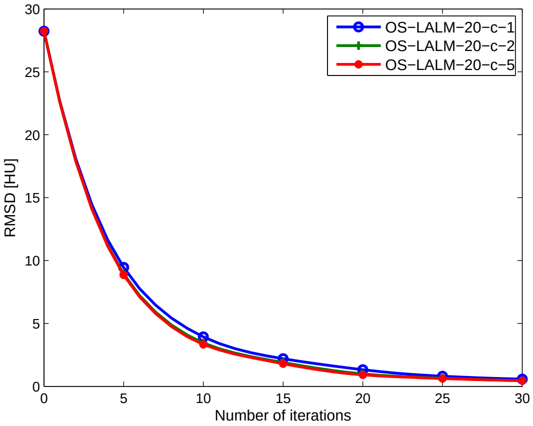

Fig. 5: Shoulder scan: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration using the proposed algorithm with different number of FISTA iterations n (1, 2, and 5) for solving the inner constrained denoising problem.

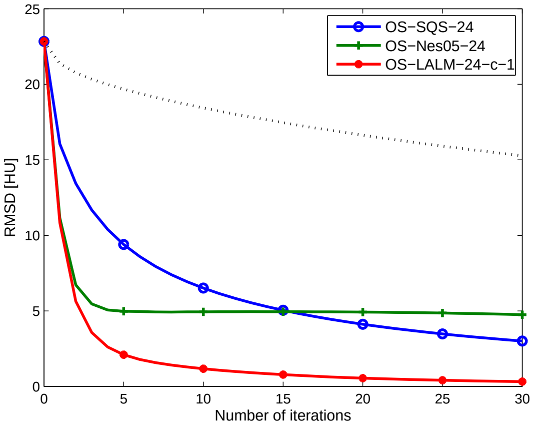

Fig. 8: GE performance phantom: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration using OS-based algorithms with 24 subsets. The dotted line shows the RMS differences using the standard OS algorithm with one subset.

Figure 7 and Figure 8 show the difference images and convergence rate curves, respectively. Because of the lower view-redundancy in axial CT scans, the OS+momentum algorithm shows even more OS artifacts than the standard OS algorithm in the difference images, leading to a larger limit cycle in Figure 8. A relaxed OS+momentum algorithm [40] that uses a diminishing step size can address this problem but require careful tuning of the step sizes. In contrast, the proposed OS-LALM algorithm avoids the need for such parameter tuning; one only needs to choose the number of subsets M.

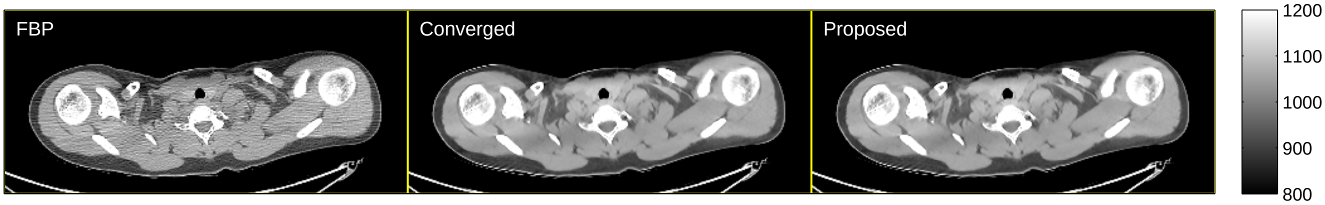

Fig. 2: Shoulder scan: cropped images (displayed from 800 to 1200 HU) from the central transaxial plane of the initial FBP image $\mathbf{x}^{(0)}$ (left), the reference reconstruction $\mathbf{x}^{\star}$ (center), and the reconstructed image using the proposed algorithm (OS-LALM-40-c-1) at the 30th iteration $\mathbf{x}^{(30)}$ (right).

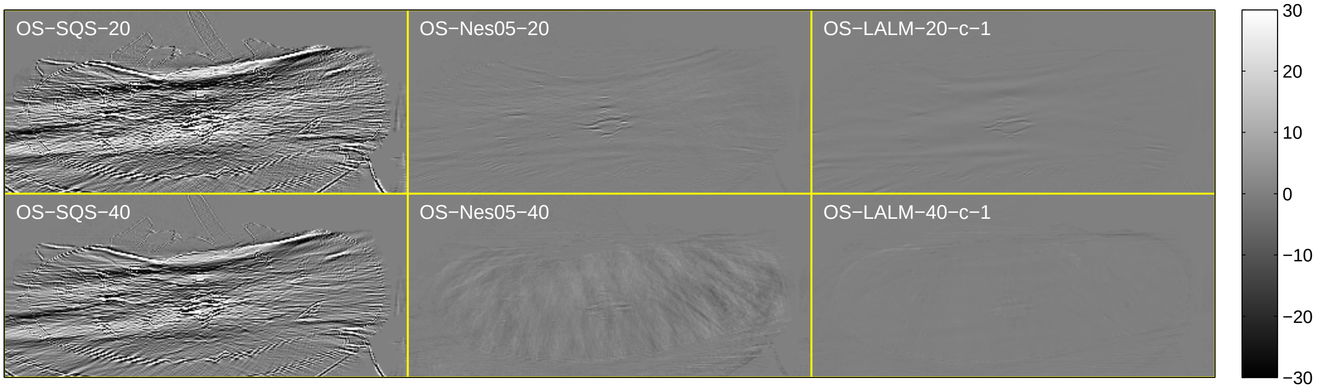

Fig. 3: Shoulder scan: cropped difference images (displayed from -30 to 30 HU) from the central transaxial plane of $\mathbf{x}^{(30)} - \mathbf{x}^{\star}$ using OS-based algorithms.

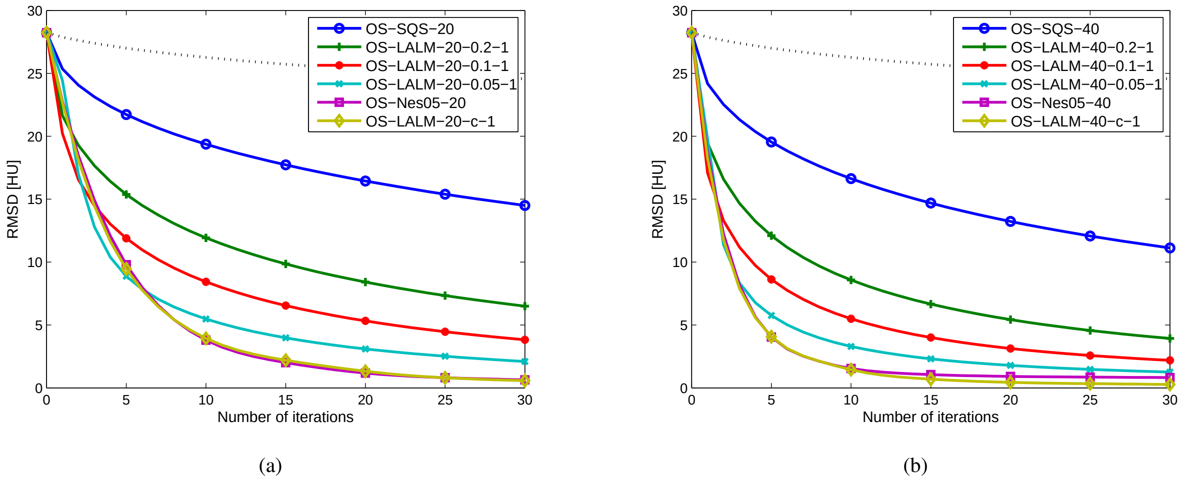

Fig. 4: Shoulder scan: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration using OS-based algorithms with (a) 20 subsets and (b) 40 subsets, respectively. The dotted lines show the RMS differences using the standard OS algorithm with one subset.

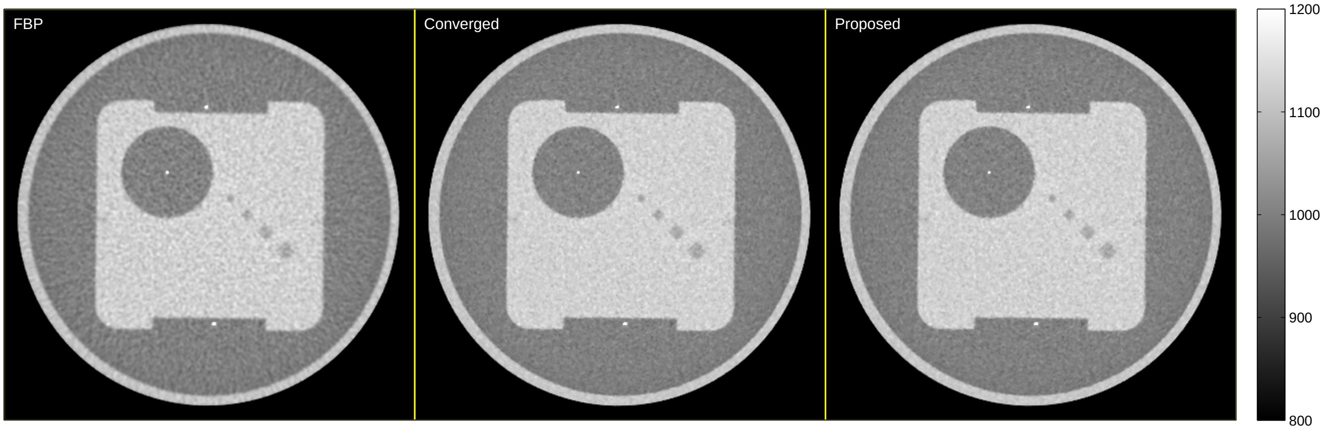

Fig. 6: GE performance phantom: cropped images (displayed from 800 to 1200 HU) from the central transaxial plane of the initial FBP image $\mathbf{x}^{(0)}$ (left), the reference reconstruction $\mathbf{x}^{\star}$ (center), and the reconstructed image using the proposed algorithm (OS-LALM-24-c-1) at the 30th iteration $\mathbf{x}^{(30)}$ (right).

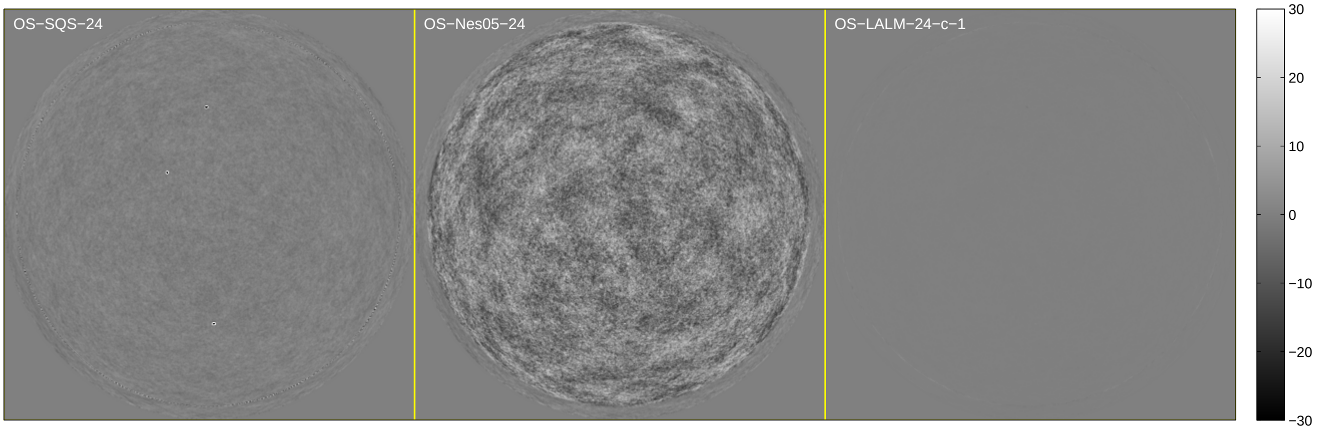

Fig. 7: GE performance phantom: cropped difference images (displayed from -30 to 30 HU) from the central transaxial plane of $\mathbf{x}^{(30)} - \mathbf{x}^{\star}$ using OS-based algorithms.

VI. Conclusion

Augmented Lagrangian (AL) methods and ordered subsets (OS) are two powerful techniques for accelerating optimization algorithms using decomposition and approximation, respectively. This paper combined these two techniques by considering a linearized variant of the AL method and proposed a fast OS-accelerable splitting-based algorithm, OS-LALM, for solving regularized (weighted) least-squares problems. We also proposed a novel deterministic downward continuation approach based on a second-order damping system that simplifies parameter selection; only the number of subsets M needs to be selected, and we provided heuristics for that based on sampling considerations in (55) and (57). We applied OS-LALM to X-ray computed tomography (CT) image reconstruction problems and compared with some state-of-the-art OS methods using real CT scans with different geometries. Experimental results showed that OS-LALM exhibits fast convergence rate and excellent gradient error tolerance.

In (33), the search direction $\mathbf{s}$ is a weighted average of the current gradient and the split gradient of L corresponding to a low-pass infinite-impulse-response filter (across iterations). The gradient error might be suppressed by this low-pass filter, improving stability. A similar averaging technique (with a low-pass finite-impulse-response filter) is used in the stochastic average gradient (SAG) method [41, 42]. In contrast, the OS+momentum algorithm computes the search direction using only the current gradient (of the auxiliary image), so the gradient error can accumulate when OS is used, providing a less stable reconstruction. When the inner constrained denoising problem is more difficult to solve, one could run more FISTA iterations or introduce an additional split variable for the regularizer as in [7] at the cost of higher memory burden, thus leading to a “high-memory” version of OS-LALM [43]. A recent alternative without variable splitting would be to use a grouped coordinate descent (GCD) denoising with a GPU implementation [44, 45].

As future work, we are interested in the convergence rate analysis of the proposed algorithm with the deterministic downward continuation approach and a more rigorous convergence analysis of OS-LALM for M that is greater than one.

Acknowledgements

The authors thank GE Healthcare for providing sinogram data in our experiments.

References

[1] M. R. Hestenes, "Multiplier and gradient methods," *J. Optim. Theory Appl.*, vol. 4, pp. 303–20, Nov. 1969.

[2] R. Glowinski and A. Marrocco, "Sur lapproximation par elements nis dordre un, et la resolution par penalisation-dualite dune classe de problemes de dirichlet nonlineaires, rev. francaise daut," *Inf. Rech. Oper.*, vol. R-2, pp. 41–76, 1975.

[3] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite-element approximations," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[4] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, Apr. 1992.

[5] T. Goldstein and S. Osher, "The split Bregman method for L1-regularized problems," *SIAM J. Imaging Sci.*, vol. 2, no. 2, pp. 323–43, 2009.

[6] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, "An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Trans. Im. Proc.*, vol. 20, pp. 681–95, Mar. 2011.

[7] S. Ramani and J. A. Fessler, "A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction," *IEEE Trans. Med. Imag.*, vol. 31, pp. 677–88, Mar. 2012.

[8] H. Erdoğan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, vol. 44, pp. 2835–51, Nov. 1999.

[9] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," 2010. August 2010 (revised December 2010) Report LIDS 2848.

[10] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Stat.*, vol. 22, pp. 400–7, Sept. 1951.

[11] D. Kim, S. Ramani, and J. A. Fessler, "Ordered subsets with momentum for accelerated X-ray CT image reconstruction," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, pp. 920–3, 2013.

[12] D. Kim, S. Ramani, and J. A. Fessler, "Accelerating X-ray CT ordered subsets image reconstruction with Nesterov's first-order methods," in *Proc. Intl. Mtg. on Fully 3D Image Recon. in Rad. and Nuc. Med*, pp. 22–5, 2013.

[13] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$," *Dokl. Akad. Nauk. USSR*, vol. 269, no. 3, pp. 543–7, 1983.

[14] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, pp. 127–52, May 2005.

[15] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[16] O. Devolder, "Stochastic first order methods in smooth convex optimization," 2011.

[17] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Adv. in Neural Info. Proc. Sys.*, pp. 1458–66, 2011.

[18] S. Ahn and J. A. Fessler, "Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms," *IEEE Trans. Med. Imag.*, vol. 22, pp. 613–26, May 2003.

[19] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero, "Convergent incremental optimization transfer algorithms: Application to tomography," *IEEE Trans. Med. Imag.*, vol. 25, pp. 283–96, Mar. 2006.

[20] H. Ouyang, N. He, L. Tran, and A. G. Gray, "Stochastic alternating direction method of multipliers," in *Proc. Intl. Conf. Machine Learning* (S. Dasgupta and D. Mcallester, eds.), vol. 28, pp. 80–8, JMLR Workshop and Conference Proceedings, 2013.

[21] Z. Lin, R. Liu, and Z. Su, "Linearized alternating direction method with adaptive penalty for low-rank representation," in *Adv. in Neural Info. Proc. Sys.*, pp. 612–20, 2011.

[22] E. Esser, X. Zhang, and T. Chan, "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM J. Imaging Sci.*, vol. 3, no. 4, pp. 1015–46, 2010.

[23] X. Zhang, M. Burger, X. Bresson, and S. Osher, "Bregmanized nonlocal regularization for deconvolution and sparse reconstruction," *SIAM J. Imaging Sci.*, vol. 3, no. 3, pp. 253–76, 2010.

[24] X. Zhang, M. Burger, and S. Osher, "A unified primal-dual algorithm framework based on Bregman iteration," *Journal of Scientific Computing*, vol. 46, no. 1, pp. 20–46, 2011.

[25] D. Kim, D. Pal, J.-B. Thibault, and J. A. Fessler, "Accelerating ordered subsets image reconstruction for X-ray CT using spatially non-uniform optimization transfer," *IEEE Trans. Med. Imag.*, vol. 32, pp. 1965–78, Nov. 2013.

[26] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Im. Vision*, vol. 40, no. 1, pp. 120–45, 2011.

[27] R. T. Rockafellar, *Convex analysis*. Princeton: Princeton University Press, 1970.

[28] I. Daubechies, M. Defrise, and C. D. Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, pp. 1413–57, Nov. 2004.

[29] D. P. Bertsekas, *Nonlinear programming*. Belmont: Athena Scientific, 2 ed., 1999.

[30] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. & Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2010.

[31] Y. Nesterov, "On an approach to the construction of optimal methods of minimization of smooth convex functions," *Ekonomika i Mateaticheskie Metody*, vol. 24, pp. 509–17, 1988. In Russian.

[32] B. O'Donoghue and E. Candès, "Adaptive restart for accelerated gradient schemes," *Found. Comput. Math.*, vol. 13, July 2013.

[33] S. Kontogiorgis and R. R. Meyer, "A variable-penalty alternating directions method for convex optimization," *Mathematical Programming*, vol. 83, no. 1–3, pp. 29–53, 1998.

[34] J. A. Fessler and W. L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs," *IEEE Trans. Im. Proc.*, vol. 5, pp. 1346–58, Sept. 1996.

[35] G. T. Herman and L. B. Meyer, "Algebraic reconstruction techniques can be made computationally efficient," *IEEE Trans. Med. Imag.*, vol. 12, pp. 600–9, Sept. 1993.

[36] J. Barzilai and J. Borwein, "Two-point step size gradient methods," *IMA J. Numerical Analysis*, vol. 8, no. 1, pp. 141–8, 1988.

[37] P. J. Huber, *Robust statistics*. New York: Wiley, 1981.

[38] J.-B. Thibault, K. Sauer, C. Bouman, and J. Hsieh, "A three-dimensional statistical approach to improved image quality for multi-slice helical CT," *Med. Phys.*, vol. 34, pp. 4526–44, Nov. 2007.

[39] W. P. Shuman, D. E. Green, J. M. Busey, O. Kolokythas, L. M. Mitsumori, K. M. Koprowicz, J.-B. Thibault, J. Hsieh, A. M. Alessio, E. Choi, and P. E. Kinahan, "Model based iterative reconstruction versus adaptive statistical iterative reconstruction and filtered back projection in 64-MDCT: Focal lesion detection, lesion conspicuity, and image noise," *Am. J. Roentgenol.*, vol. 200, pp. 1071–6, May 2013.

[40] D. Kim and J. A. Fessler, "Ordered subsets acceleration using relaxed momentum for X-ray CT image reconstruction," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, pp. 1–5, 2013.

[41] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets," in *Adv. in Neural Info. Proc. Sys.*, pp. 2672–80, 2012.

[42] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," 2013.

[43] H. Nien and J. A. Fessler, "Fast splitting-based ordered-subsets X-ray CT image reconstruction," in *Proc. 3rd Intl. Mtg. on image formation in X-ray CT*, pp. 291–4, 2014.

[44] M. G. McGaffin and J. A. Fessler, "Fast edge-preserving image denoising via group coordinate descent on the GPU," in *Proc. SPIE 9020 Computational Imaging XII*, p. 90200P, 2014.

[45] M. G. McGaffin and J. A. Fessler, "Edge-preserving image denoising via group coordinate descent on the GPU," *IEEE Trans. Im. Proc.*, 2014. Submitted.

Fast X-Ray CT Image Reconstruction Using a Linearized Augmented Lagrangian Method with Ordered Subsets: Supplementary Material

Hung Nien, *Student Member, IEEE*, and Jeffrey A. Fessler, *Fellow, IEEE*

September 11, 2014

In this supplementary material, we provide the detailed convergence analysis of a linearized augmented Lagrangian (AL) method with inexact updates proposed in [1] together with additional experimental results.

I. Convergence analysis of the inexact linearized AL method

Consider a general composite convex optimization problem:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \left\{ g(\mathbf{A}\mathbf{x}) + h(\mathbf{x}) \right\} \tag{1}$$

and its equivalent constrained minimization problem:

$$(\hat{\mathbf{x}}, \hat{\mathbf{u}}) \in \arg\min_{\mathbf{x},\mathbf{u}} \left\{ g(\mathbf{u}) + h(\mathbf{x}) \right\} \text{ s.t. } \mathbf{u} = \mathbf{A}\mathbf{x}\,, \tag{2}$$

where both g and h are closed and proper convex functions. The two inexact linearized AL method (LALM) variants that solve (2) are as follows:

$$\begin{cases} \left\| \mathbf{x}^{(k+1)} - \arg\min_{\mathbf{x}} \phi_k(\mathbf{x}) \right\| \leq \delta_k \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)}\,, \end{cases} \tag{3}$$

and

$$\begin{cases} \left| \phi_k\left(\mathbf{x}^{(k+1)}\right) - \min_{\mathbf{x}} \phi_k(\mathbf{x}) \right| \leq \varepsilon_k \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)}\,, \end{cases} \tag{4}$$

where

$$\phi_k(\mathbf{x}) \triangleq h(\mathbf{x}) + \breve{\theta}_k\left(\mathbf{x}; \mathbf{x}^{(k)}\right)\,, \tag{5}$$

and

$$\breve{\theta}_k\left(\mathbf{x}; \mathbf{x}^{(k)}\right) \triangleq \theta_k\left(\mathbf{x}^{(k)}\right) + \left\langle \nabla\theta_k\left(\mathbf{x}^{(k)}\right), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle + \frac{\rho L}{2} \left\| \mathbf{x} - \mathbf{x}^{(k)} \right\|_2^2 \tag{6}$$

is the separable quadratic surrogate (SQS) function of

$$\theta_k(\mathbf{x}) \triangleq \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)} \right\|_2^2 \tag{7}$$

with $L > \|\mathbf{A}\|_2^2 = \lambda_{\max}(\mathbf{A}'\mathbf{A})$, $\{\delta_k\}_{k=0}^{\infty}$ and $\{\varepsilon_k\}_{k=0}^{\infty}$ are two non-negative sequences, $\mathbf{d}$ is the scaled Lagrange multiplier of the split variable $\mathbf{u}$, and $\rho > 0$ is the corresponding AL penalty parameter. Furthermore, in [1], we also showed that the inexact LALM (with $\mathbf{u}^{(0)} = \mathbf{A}\mathbf{x}^{(0)}$) is equivalent to the invexact version of the Chambolle-Pock first-order primal-dual algorithm (CP) [2]:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \operatorname{prox}_{\sigma h}\left(\mathbf{x}^{(k)} - \sigma\mathbf{A}'\bar{\mathbf{z}}^{(k)}\right) \\ \mathbf{z}^{(k+1)} \in \operatorname{prox}_{\tau g^*}\left(\mathbf{z}^{(k)} + \tau\mathbf{A}\mathbf{x}^{(k+1)}\right) \\ \bar{\mathbf{z}}^{(k+1)} = \mathbf{z}^{(k+1)} + \left(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\right) \end{cases} \tag{8}$$

that solves the minimax problem:

$$(\hat{\mathbf{z}}, \hat{\mathbf{x}}) \in \arg\min_{\mathbf{z}} \max_{\mathbf{x}} \left\{ \Omega(\mathbf{z}, \mathbf{x}) \triangleq \langle -\mathbf{A}'\mathbf{z}, \mathbf{x} \rangle + g^*(\mathbf{z}) - h(\mathbf{x}) \right\} \tag{9}$$

This work is supported in part by National Institutes of Health (NIH) grants R01-HL-098686 and U01-EB-18753 and by an equipment donation from Intel Corporation. Hung Nien and Jeffrey A. Fessler are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48105, USA (e-mail: {hungnien, fessler}@umich.edu).

with $\mathbf{z} = -\tau\mathbf{d}$, $\bar{\mathbf{z}}^{(0)} = \mathbf{z}^{(0)}$, $\sigma = \rho^{-1}t$, $\tau = \rho$, and $t \triangleq 1/L$, where prox_f denotes the proximal mapping of f defined as:

$$\operatorname{prox}_f(\mathbf{z}) \triangleq \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \tfrac{1}{2} \|\mathbf{x} - \mathbf{z}\|_2^2 \right\} , \tag{10}$$

and f^* denotes the convex conjugate of a function f. Note that $g^{**} = g$ and $h^{**} = h$ since both g and h are closed, proper, and convex.

A. Proof of Theorem 1

Theorem 1. *Consider a constrained composite convex optimization problem (2) where both g and h are closed and proper convex functions. Let $\rho > 0$ and $\{\delta_k\}_{k=0}^{\infty}$ denote a non-negative sequence such that*

$$\sum_{k=0}^{\infty} \delta_k < \infty . \tag{11}$$

If (2) has a solution $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$, then the sequence of updates $\left\{\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}\right)\right\}_{k=0}^{\infty}$ generated by the inexact LALM in (3) converges to $(\hat{\mathbf{x}}, \hat{\mathbf{u}})$; otherwise, at least one of the sequences $\left\{\left(\mathbf{x}^{(k)}, \mathbf{u}^{(k)}\right)\right\}_{k=0}^{\infty}$ or $\left\{\mathbf{d}^{(k)}\right\}_{k=0}^{\infty}$ diverges.

Proof. To prove this theorem, we first consider the exact LALM:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \breve{\theta}_k\left(\mathbf{x}; \mathbf{x}^{(k)}\right) \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \tfrac{\rho}{2} \left\|\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)}\right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} . \end{cases} \tag{12}$$

Note that

$$\begin{aligned} \breve{\theta}_k\left(\mathbf{x}; \mathbf{x}^{(k)}\right) &= \theta_k\left(\mathbf{x}^{(k)}\right) + \left\langle \nabla\theta_k\left(\mathbf{x}^{(k)}\right), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle + \tfrac{\rho L}{2} \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_2^2 \\ &= \theta_k\left(\mathbf{x}^{(k)}\right) + \left\langle \nabla\theta_k\left(\mathbf{x}^{(k)}\right), \mathbf{x} - \mathbf{x}^{(k)} \right\rangle + \tfrac{\rho}{2} \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_{\mathbf{A}'\mathbf{A}}^2 + \tfrac{\rho}{2} \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_{L\mathbf{I}-\mathbf{A}'\mathbf{A}}^2 \\ &= \theta_k(\mathbf{x}) + \tfrac{\rho}{2} \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_{\mathbf{G}}^2 , \end{aligned} \tag{13}$$

where $\mathbf{G} \triangleq L\mathbf{I} - \mathbf{A}'\mathbf{A} \succ 0$. Therefore, the exact LALM can also be written as

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \tfrac{\rho}{2} \left\|\mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)}\right\|_2^2 + \tfrac{\rho}{2} \left\|\mathbf{x} - \mathbf{x}^{(k)}\right\|_{\mathbf{G}}^2 \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \tfrac{\rho}{2} \left\|\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)}\right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} . \end{cases} \tag{14}$$

Now, consider another constrained minimization problem that is also equivalent to (1) but uses two split variables:

$$(\hat{\mathbf{x}}, \hat{\mathbf{u}}, \hat{\mathbf{v}}) \in \arg\min_{\mathbf{x},\mathbf{u},\mathbf{v}} \left\{ g(\mathbf{u}) + h(\mathbf{x}) \right\} \text{ s.t. } \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{A} \\ \mathbf{G}^{1/2} \end{bmatrix}}_{\mathbf{S}} \mathbf{x} . \tag{15}$$

The corresponding augmented Lagrangian and ADMM iterates [3] are

$$\mathcal{L}_{\mathrm{A}}(\mathbf{x}, \mathbf{u}, \mathbf{d}, \mathbf{v}, \mathbf{e}; \rho, \eta) \triangleq g(\mathbf{u}) + h(\mathbf{x}) + \tfrac{\rho}{2} \|\mathbf{A}\mathbf{x} - \mathbf{u} - \mathbf{d}\|_2^2 + \tfrac{\eta}{2} \left\|\mathbf{G}^{1/2}\mathbf{x} - \mathbf{v} - \mathbf{e}\right\|_2^2 \tag{16}$$

and

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \tfrac{\rho}{2} \left\|\mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)}\right\|_2^2 + \tfrac{\eta}{2} \left\|\mathbf{G}^{1/2}\mathbf{x} - \mathbf{v}^{(k)} - \mathbf{e}^{(k)}\right\|_2^2 \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \tfrac{\rho}{2} \left\|\mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)}\right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} \\ \mathbf{v}^{(k+1)} = \mathbf{G}^{1/2}\mathbf{x}^{(k+1)} - \mathbf{e}^{(k)} \\ \mathbf{e}^{(k+1)} = \mathbf{e}^{(k)} - \mathbf{G}^{1/2}\mathbf{x}^{(k+1)} + \mathbf{v}^{(k+1)} , \end{cases} \tag{17}$$

where $\mathbf{e}$ is the scaled Lagrange multiplier of the split variable $\mathbf{v}$, and $\eta > 0$ is the corresponding AL penalty parameter. Note that since $\mathbf{G}$ is positive definite, $\mathbf{S}$ defined in (15) has full column rank. Hence, the ADMM iterates (17) are convergent [4, Theorem 8]. Solving the last two iterates in (17) yields identities

$$\begin{cases} \mathbf{v}^{(k+1)} = \mathbf{G}^{1/2}\mathbf{x}^{(k+1)} \\ \mathbf{e}^{(k+1)} = \mathbf{0} \end{cases} \tag{18}$$

if we initialize $\mathbf{e}$ as $\mathbf{e}^{(0)} = \mathbf{0}$. Substituting (18) into (17), we have the equivalent ADMM iterates:

$$\begin{cases} \mathbf{x}^{(k+1)} \in \arg\min_{\mathbf{x}} \left\{ h(\mathbf{x}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x} - \mathbf{u}^{(k)} - \mathbf{d}^{(k)} \right\|_2^2 + \frac{\eta}{2} \left\| \mathbf{G}^{1/2}\mathbf{x} - \mathbf{G}^{1/2}\mathbf{x}^{(k)} \right\|_2^2 \right\} \\ \mathbf{u}^{(k+1)} \in \arg\min_{\mathbf{u}} \left\{ g(\mathbf{u}) + \frac{\rho}{2} \left\| \mathbf{A}\mathbf{x}^{(k+1)} - \mathbf{u} - \mathbf{d}^{(k)} \right\|_2^2 \right\} \\ \mathbf{d}^{(k+1)} = \mathbf{d}^{(k)} - \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{u}^{(k+1)} . \end{cases} \tag{19}$$

When $\eta = \rho$, the equivalent ADMM iterates (19) reduce to (14). Therefore, LALM is a convergent ADMM. Finally, by using [4, Theorem 8], LALM is convergent if the error of $\mathbf{x}$-update is summable. That is, the inexact LALM in (3) is convergent if the non-negative sequence $\{\delta_k\}_{k=0}^{\infty}$ satisfies $\sum_{k=0}^{\infty} \delta_k < \infty$. ∎

B. Proof of Theorem 2

Theorem 2. *Consider a minimax problem (9) where both g and h are closed and proper convex functions. Suppose it has a saddle-point $(\hat{\mathbf{z}}, \hat{\mathbf{x}})$. Note that since the minimization problem (1) happens to be the dual problem of (9), $\hat{\mathbf{x}}$ is also a solution of (1). Let $\rho > 0$ and $\{\varepsilon_k\}_{k=0}^{\infty}$ denote a non-negative sequence such that*

$$\sum_{k=0}^{\infty} \sqrt{\varepsilon_k} < \infty . \tag{20}$$

Then, the sequence of updates $\left\{ \left(-\rho\mathbf{d}^{(k)}, \mathbf{x}^{(k)}\right) \right\}_{k=0}^{\infty}$ generated by the inexact LALM in (4) is a bounded sequence that converges to $(\hat{\mathbf{z}}, \hat{\mathbf{x}})$, and the primal-dual gap of $(\mathbf{z}_k, \mathbf{x}_k)$ has the following bound:

$$\Omega\left(\mathbf{z}_k, \hat{\mathbf{x}}\right) - \Omega\left(\hat{\mathbf{z}}, \mathbf{x}_k\right) \leq \frac{\left(C + 2A_k + \sqrt{B_k}\right)^2}{k} , \tag{21}$$

where $\mathbf{z}_k \triangleq \frac{1}{k} \sum_{j=1}^{k} \left(-\rho\mathbf{d}^{(j)}\right)$, $\mathbf{x}_k \triangleq \frac{1}{k} \sum_{j=1}^{k} \mathbf{x}^{(j)}$,

$$C \triangleq \frac{\left\| \mathbf{x}^{(0)} - \hat{\mathbf{x}} \right\|_2}{\sqrt{2\rho^{-1}t}} + \frac{\left\| \left(-\rho\mathbf{d}^{(0)}\right) - \hat{\mathbf{z}} \right\|_2}{\sqrt{2\rho}} , \tag{22}$$

$$A_k \triangleq \sum_{j=1}^{k} \sqrt{\frac{\varepsilon_{j-1}}{\left(1 - t \left\| \mathbf{A} \right\|_2^2\right) \rho^{-1} t}} , \tag{23}$$

and

$$B_k \triangleq \sum_{j=1}^{k} \varepsilon_{j-1} . \tag{24}$$

Proof. As mentioned before, the inexact LALM is the inexact version of CP with a specific choice of σ and τ and a substitution $\mathbf{z} = -\tau\mathbf{d}$ (if we initialize both algorithms appropriately). Here, we just prove the convergence of the inexact CP by extending the analysis in [2], and the inexact LALM is simply a special case of the inexact CP. However, since the proximal mapping in the $\mathbf{x}$-update of the inexact CP is solved inexactly, the existing analysis is not applicable. To solve this problem, we adopt the error analysis technique developed in [5]. We first define the inexact proximal mapping

$$\mathbf{u} \overset{\varepsilon}{\approx} \operatorname{prox}_{\phi}(\mathbf{v}) \tag{25}$$

to be the mapping that satisfies

$$\phi(\mathbf{u}) + \tfrac{1}{2} \left\| \mathbf{u} - \mathbf{v} \right\|_2^2 \leq \varepsilon + \min_{\bar{\mathbf{u}}} \left\{ \phi(\bar{\mathbf{u}}) + \tfrac{1}{2} \left\| \bar{\mathbf{u}} - \mathbf{v} \right\|_2^2 \right\} . \tag{26}$$

Therefore, the inexact CP is defined as

$$\begin{cases} \mathbf{x}^{(k+1)} \overset{\varepsilon_k}{\approx} \operatorname{prox}_{\sigma h}\left(\mathbf{x}^{(k)} - \sigma\mathbf{A}'\bar{\mathbf{z}}^{(k)}\right) \\ \mathbf{z}^{(k+1)} \in \operatorname{prox}_{\tau g^*}\left(\mathbf{z}^{(k)} + \tau\mathbf{A}\mathbf{x}^{(k+1)}\right) \\ \bar{\mathbf{z}}^{(k+1)} = \mathbf{z}^{(k+1)} + \left(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}\right) \end{cases} \tag{27}$$

with $\sigma\tau \left\| \mathbf{A} \right\|_2^2 < 1$. One can verify that with $\mathbf{z} = -\tau\mathbf{d}$, $\sigma = \rho^{-1}t$, and $\tau = \rho$, the inexact CP in (27) is equivalent to the inexact LALM in (4). Schmidt *et al.* showed that

$$\mathbf{u} \overset{\varepsilon}{\approx} \operatorname{prox}_{\phi}(\mathbf{v}) \Leftrightarrow \mathbf{v} - \mathbf{u} - \mathbf{f} \in \partial_{\varepsilon}\phi(\mathbf{u}) \tag{28}$$

with $\left\| \mathbf{f} \right\|_2 \leq \sqrt{2\varepsilon}$, and for any $\mathbf{s} \in \partial_{\varepsilon}\phi(\mathbf{u})$,

$$\phi(\mathbf{w}) \geq \phi(\mathbf{u}) + \mathbf{s}'\left(\mathbf{w} - \mathbf{u}\right) - \varepsilon \tag{29}$$

for all $\mathbf{w}$, where $\partial_\varepsilon \phi(\mathbf{u})$ denotes the ε-subdifferential of ϕ at $\mathbf{u}$ [5, Lemma 2]. When $\varepsilon = 0$, (28) and (29) reduce to the standard optimality condition of a proximal mapping and the definition of subgradient, respectively. At the jth iteration, $j = 0, \ldots, k-1$, the updates generated by the inexact CP in (27) satisfy

$$\begin{cases} \left(\mathbf{x}^{(j)} - \sigma \mathbf{A}' \bar{\mathbf{z}}^{(j)}\right) - \mathbf{x}^{(j+1)} - \mathbf{f}^{(j)} \in \partial_{\varepsilon_j} \left(\sigma h\right) \left(\mathbf{x}^{(j+1)}\right) \\ \left(\mathbf{z}^{(j)} + \tau \mathbf{A}\mathbf{x}^{(j+1)}\right) - \mathbf{z}^{(j+1)} \in \partial \left(\tau g^*\right) \left(\mathbf{z}^{(j+1)}\right). \end{cases} \tag{30}$$

In other words,

$$\frac{\mathbf{x}^{(j)} - \mathbf{x}^{(j+1)}}{\sigma} - \mathbf{A}'\bar{\mathbf{z}}^{(j)} - \frac{\mathbf{f}^{(j)}}{\sigma} \in \partial_{\varepsilon_j} h\big(\mathbf{x}^{(j+1)}\big) \tag{31}$$

and

$$\frac{\mathbf{z}^{(j)} - \mathbf{z}^{(j+1)}}{\tau} + \mathbf{A}\mathbf{x}^{(j+1)} \in \partial g^*\big(\mathbf{z}^{(j+1)}\big)\,, \tag{32}$$

where $\big\|\mathbf{f}^{(j)}\big\|_2 \le \sqrt{2\varepsilon_j}$. From (31), we have

$$\begin{aligned} h(\mathbf{x}) &\ge h\big(\mathbf{x}^{(j+1)}\big) + \big\langle \partial_{\varepsilon_j} h\big(\mathbf{x}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \varepsilon_j \\ &= h\big(\mathbf{x}^{(j+1)}\big) + \big\langle \tfrac{\mathbf{x}^{(j)} - \mathbf{x}^{(j+1)}}{\sigma}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \big\langle \mathbf{A}'\bar{\mathbf{z}}^{(j)}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \big\langle \tfrac{\mathbf{f}^{(j)}}{\sigma}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \varepsilon_j \\ &= h\big(\mathbf{x}^{(j+1)}\big) + \tfrac{1}{2\sigma}\big(\big\|\mathbf{x}^{(j+1)} - \mathbf{x}\big\|_2^2 + \big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2 - \big\|\mathbf{x}^{(j)} - \mathbf{x}\big\|_2^2 \big) \\ &\qquad + \big\langle -\mathbf{A}'\big(\bar{\mathbf{z}}^{(j)} - \mathbf{z}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle + \big\langle -\mathbf{A}'\mathbf{z}^{(j+1)}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \big\langle \tfrac{\mathbf{f}^{(j)}}{\sigma}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \varepsilon_j \\ &\ge h\big(\mathbf{x}^{(j+1)}\big) + \tfrac{1}{2\sigma}\big(\big\|\mathbf{x}^{(j+1)} - \mathbf{x}\big\|_2^2 + \big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2 - \big\|\mathbf{x}^{(j)} - \mathbf{x}\big\|_2^2 \big) \\ &\qquad + \big\langle -\mathbf{A}'\big(\bar{\mathbf{z}}^{(j)} - \mathbf{z}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle + \big\langle -\mathbf{A}'\mathbf{z}^{(j+1)}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \tfrac{1}{\sigma}\,\big\|\mathbf{f}^{(j)}\big\|_2 \big\|\mathbf{x} - \mathbf{x}^{(j+1)}\big\|_2 - \varepsilon_j \\ &\ge h\big(\mathbf{x}^{(j+1)}\big) + \tfrac{1}{2\sigma}\big(\big\|\mathbf{x}^{(j+1)} - \mathbf{x}\big\|_2^2 + \big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2 - \big\|\mathbf{x}^{(j)} - \mathbf{x}\big\|_2^2 \big) \\ &\qquad + \big\langle -\mathbf{A}'\big(\bar{\mathbf{z}}^{(j)} - \mathbf{z}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle + \big\langle -\mathbf{A}'\mathbf{z}^{(j+1)}, \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \tfrac{\sqrt{2\varepsilon_j}}{\sigma}\,\big\|\mathbf{x} - \mathbf{x}^{(j+1)}\big\|_2 - \varepsilon_j \end{aligned} \tag{33}$$

for any $\mathbf{x} \in \operatorname{Dom} h$. From (32), we have

$$\begin{aligned} g^*(\mathbf{z}) &\ge g^*\big(\mathbf{z}^{(j+1)}\big) + \big\langle \partial g^*\big(\mathbf{z}^{(j+1)}\big), \mathbf{z} - \mathbf{z}^{(j+1)}\big\rangle \\ &= g^*\big(\mathbf{z}^{(j+1)}\big) + \big\langle \tfrac{\mathbf{z}^{(j)} - \mathbf{z}^{(j+1)}}{\tau}, \mathbf{z} - \mathbf{z}^{(j+1)}\big\rangle + \big\langle \mathbf{A}\mathbf{x}^{(j+1)}, \mathbf{z} - \mathbf{z}^{(j+1)}\big\rangle \\ &= g^*\big(\mathbf{z}^{(j+1)}\big) + \tfrac{1}{2\tau}\big(\big\|\mathbf{z}^{(j+1)} - \mathbf{z}\big\|_2^2 + \big\|\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big\|_2^2 - \big\|\mathbf{z}^{(j)} - \mathbf{z}\big\|_2^2 \big) - \big\langle -\mathbf{A}'\big(\mathbf{z} - \mathbf{z}^{(j+1)}\big), \mathbf{x}^{(j+1)}\big\rangle \end{aligned} \tag{34}$$

for any $\mathbf{z} \in \operatorname{Dom} g^*$. Summing (33) and (34), it follows:

$$\begin{aligned} \frac{\big\|\mathbf{x}^{(j)} - \mathbf{x}\big\|_2^2}{2\sigma} + \frac{\big\|\mathbf{z}^{(j)} - \mathbf{z}\big\|_2^2}{2\tau} &\ge \Big(\Omega\big(\mathbf{z}^{(j+1)}, \mathbf{x}\big) - \Omega\big(\mathbf{z}, \mathbf{x}^{(j+1)}\big) \Big) \\ &\quad + \frac{\big\|\mathbf{x}^{(j+1)} - \mathbf{x}\big\|_2^2}{2\sigma} + \frac{\big\|\mathbf{z}^{(j+1)} - \mathbf{z}\big\|_2^2}{2\tau} + \frac{\big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2}{2\sigma} + \frac{\big\|\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big\|_2^2}{2\tau} \\ &\quad + \big\langle -\mathbf{A}'\big(\bar{\mathbf{z}}^{(j)} - \mathbf{z}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle - \frac{\sqrt{2\varepsilon_j}}{\sigma}\,\big\|\mathbf{x} - \mathbf{x}^{(j+1)}\big\|_2 - \varepsilon_j\,. \end{aligned} \tag{35}$$

Furthermore,

$$\begin{aligned} &\big\langle -\mathbf{A}'\big(\bar{\mathbf{z}}^{(j)} - \mathbf{z}^{(j+1)}\big), \mathbf{x} - \mathbf{x}^{(j+1)}\big\rangle \\ &= \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j+1)} - 2\mathbf{z}^{(j)} + \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}\big\rangle \\ &= \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}\big\rangle - \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j)} - \mathbf{x}\big\rangle - \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\rangle \\ &\ge \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}\big\rangle - \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j)} - \mathbf{x}\big\rangle - \|\mathbf{A}\|_2 \big\|\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big\|_2 \big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2 \\ &\ge \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}\big\rangle - \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j)} - \mathbf{x}\big\rangle \\ &\qquad\qquad - \|\mathbf{A}\|_2 \big(\tfrac{\sqrt{\sigma/\tau}}{2} \big\|\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big\|_2^2 + \tfrac{1}{2\sqrt{\sigma/\tau}} \big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2 \big) \qquad (36) \\ &\ge \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j+1)} - \mathbf{z}^{(j)}\big), \mathbf{x}^{(j+1)} - \mathbf{x}\big\rangle - \big\langle -\mathbf{A}'\big(\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big), \mathbf{x}^{(j)} - \mathbf{x}\big\rangle \\ &\qquad\qquad - \sqrt{\sigma\tau}\,\|\mathbf{A}\|_2 \left(\frac{\big\|\mathbf{z}^{(j)} - \mathbf{z}^{(j-1)}\big\|_2^2}{2\tau} + \frac{\big\|\mathbf{x}^{(j+1)} - \mathbf{x}^{(j)}\big\|_2^2}{2\sigma} \right), \qquad (37) \end{aligned}$$

where (36) is due to Young's inequality. Plugging (37) into (35), it follows that for any $(\mathbf{z},\mathbf{x})$,

$$\begin{aligned}\frac{\left\|\mathbf{x}^{(j)}-\mathbf{x}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(j)}-\mathbf{z}\right\|_2^2}{2\tau} &\geq \Big(\Omega\big(\mathbf{z}^{(j+1)},\mathbf{x}\big)-\Omega\big(\mathbf{z},\mathbf{x}^{(j+1)}\big)\Big)+\frac{\left\|\mathbf{x}^{(j+1)}-\mathbf{x}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(j+1)}-\mathbf{z}\right\|_2^2}{2\tau}\\ &+\big(1-\sqrt{\sigma\tau}\left\|\mathbf{A}\right\|_2\big)\frac{\left\|\mathbf{x}^{(j+1)}-\mathbf{x}^{(j)}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(j+1)}-\mathbf{z}^{(j)}\right\|_2^2}{2\tau}-\sqrt{\sigma\tau}\left\|\mathbf{A}\right\|_2\frac{\left\|\mathbf{z}^{(j)}-\mathbf{z}^{(j-1)}\right\|_2^2}{2\tau}\\ &+\big\langle-\mathbf{A}'\big(\mathbf{z}^{(j+1)}-\mathbf{z}^{(j)}\big),\mathbf{x}^{(j+1)}-\mathbf{x}\big\rangle-\big\langle-\mathbf{A}'\big(\mathbf{z}^{(j)}-\mathbf{z}^{(j-1)}\big),\mathbf{x}^{(j)}-\mathbf{x}\big\rangle-\frac{\sqrt{2\varepsilon_j}}{\sigma}\left\|\mathbf{x}-\mathbf{x}^{(j+1)}\right\|_2-\varepsilon_j\,. \end{aligned}\tag{38}$$

Suppose $\mathbf{z}^{(-1)}=\mathbf{z}^{(0)}$, i.e., $\bar{\mathbf{z}}^{(0)}=\mathbf{z}^{(0)}$. Summing up (38) from $j=0,\ldots,k-1$ and using

$$\big\langle-\mathbf{A}'\big(\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\big),\mathbf{x}^{(k)}-\mathbf{x}\big\rangle\leq\frac{\left\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\right\|_2^2}{2\tau}+\sigma\tau\left\|\mathbf{A}\right\|_2^2\frac{\left\|\mathbf{x}^{(k)}-\mathbf{x}\right\|_2^2}{2\sigma}\tag{39}$$

as before, we have

$$\begin{aligned}&\sum_{j=1}^{k}\Big(\Omega\big(\mathbf{z}^{(j)},\mathbf{x}\big)-\Omega\big(\mathbf{z},\mathbf{x}^{(j)}\big)\Big)+\Big(1-\sigma\tau\left\|\mathbf{A}\right\|_2^2\Big)\frac{\left\|\mathbf{x}^{(k)}-\mathbf{x}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(k)}-\mathbf{z}\right\|_2^2}{2\tau}\\ &+\big(1-\sqrt{\sigma\tau}\left\|\mathbf{A}\right\|_2\big)\sum_{j=1}^{k}\frac{\left\|\mathbf{x}^{(j)}-\mathbf{x}^{(j-1)}\right\|_2^2}{2\sigma}+\big(1-\sqrt{\sigma\tau}\left\|\mathbf{A}\right\|_2\big)\sum_{j=1}^{k-1}\frac{\left\|\mathbf{z}^{(j)}-\mathbf{z}^{(j-1)}\right\|_2^2}{2\tau}\\ &\qquad\leq\frac{\left\|\mathbf{x}^{(0)}-\mathbf{x}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(0)}-\mathbf{z}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\varepsilon_{j-1}+\sum_{j=1}^{k}2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\frac{\left\|\mathbf{x}^{(j)}-\mathbf{x}\right\|_2}{\sqrt{2\sigma}}\,.\end{aligned}\tag{40}$$

Since $\sigma\tau\left\|\mathbf{A}\right\|_2^2<1$, we have $1-\sigma\tau\left\|\mathbf{A}\right\|_2^2>0$ and $1-\sqrt{\sigma\tau}\left\|\mathbf{A}\right\|_2>0$. If we choose $(\mathbf{z},\mathbf{x})=(\hat{\mathbf{z}},\hat{\mathbf{x}})$, the first term on the left-hand side of (40) is the sum of k non-negative primal-dual gaps, and all terms in (40) are greater than or equal to zero. Let $D\triangleq 1-\sigma\tau\left\|\mathbf{A}\right\|_2^2>0$. We have three inequalities:

$$D\cdot\frac{\left\|\mathbf{x}^{(k)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}\leq\frac{\left\|\mathbf{x}^{(0)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(0)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\varepsilon_{j-1}+\sum_{j=1}^{k}2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\frac{\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\,,\tag{41}$$

$$D\cdot\left(\frac{\left\|\mathbf{x}^{(k)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(k)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}\right)\leq\frac{\left\|\mathbf{x}^{(0)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(0)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\varepsilon_{j-1}+\sum_{j=1}^{k}2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\frac{\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\,,\tag{42}$$

and

$$\sum_{j=1}^{k}\Big(\Omega\big(\mathbf{z}^{(j)},\hat{\mathbf{x}}\big)-\Omega\big(\hat{\mathbf{z}},\mathbf{x}^{(j)}\big)\Big)\leq\frac{\left\|\mathbf{x}^{(0)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{\left\|\mathbf{z}^{(0)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\varepsilon_{j-1}+\sum_{j=1}^{k}2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\frac{\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\,.\tag{43}$$

All these inequality has a common right-hand-side. To continue the proof, we have to bound $\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2/\sqrt{2\sigma}$ first. Dividing D from both sides of (41), we have

$$\left(\frac{\left\|\mathbf{x}^{(k)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\right)^2\leq\left(\frac{1}{D}\frac{\left\|\mathbf{x}^{(0)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{1}{D}\frac{\left\|\mathbf{z}^{(0)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\frac{\varepsilon_{j-1}}{D}\right)+\sum_{j=1}^{k}2\left(\frac{1}{D}\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\right)\frac{\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\,.\tag{44}$$

Let

$$S_k\triangleq\frac{1}{D}\frac{\left\|\mathbf{x}^{(0)}-\hat{\mathbf{x}}\right\|_2^2}{2\sigma}+\frac{1}{D}\frac{\left\|\mathbf{z}^{(0)}-\hat{\mathbf{z}}\right\|_2^2}{2\tau}+\sum_{j=1}^{k}\frac{\varepsilon_{j-1}}{D}\,,\tag{45}$$

$$\lambda_j\triangleq 2\left(\frac{1}{D}\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\right)\,,\tag{46}$$

and

$$u_j\triangleq\frac{\left\|\mathbf{x}^{(j)}-\hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}}\,.\tag{47}$$

We have $u_k^2\leq S_k+\sum_{j=1}^{k}\lambda_j u_j$ from (44) with $\{S_k\}_{k=0}^{\infty}$ an increasing sequence, $S_0\geq u_0^2$ (note that $0<D<1$ because

$0 < \sigma\tau \left\|\mathbf{A}\right\|_2^2 < 1$), and $\lambda_j \geq 0$ for all j. According to [5, Lemma 1], it follows that

$$\frac{\left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} \leq \widetilde{A}_k + \left(\frac{1}{D}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{1}{D}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \widetilde{B}_k + \widetilde{A}_k^2\right)^{1/2}, \tag{48}$$

where

$$\widetilde{A}_k \triangleq \sum_{j=1}^{k} \frac{1}{D}\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\,, \tag{49}$$

and

$$\widetilde{B}_k \triangleq \sum_{j=1}^{k} \frac{\varepsilon_{j-1}}{D}\,. \tag{50}$$

Since $\widetilde{A}_j$ and $\widetilde{B}_j$ are increasing sequences of j, for $j \leq k$, we have

$$\begin{aligned}
\frac{\left\|\mathbf{x}^{(j)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} &\leq \widetilde{A}_j + \left(\frac{1}{D}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{1}{D}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \widetilde{B}_j + \widetilde{A}_j^2\right)^{1/2} \\
&\leq \widetilde{A}_k + \left(\frac{1}{D}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{1}{D}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \widetilde{B}_k + \widetilde{A}_k^2\right)^{1/2} \\
&\leq \widetilde{A}_k + \left(\frac{1}{\sqrt{D}}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{1}{\sqrt{D}}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + \sqrt{\widetilde{B}_k} + \widetilde{A}_k\right).
\end{aligned} \tag{51}$$

Now, we can bound the right-hand-side of (41), (42), and (43) as

$$\begin{aligned}
&\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \sum_{j=1}^{k}\varepsilon_{j-1} + \sum_{j=1}^{k} 2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\frac{\left\|\mathbf{x}^{(j)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} \\
&\leq \frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \sum_{j=1}^{k}\varepsilon_{j-1} + \sum_{j=1}^{k} 2\sqrt{\tfrac{\varepsilon_{j-1}}{\sigma}}\left(2\widetilde{A}_k + \frac{1}{\sqrt{D}}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{1}{\sqrt{D}}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + \sqrt{\widetilde{B}_k}\right) \\
&= \frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} + \widetilde{B}_k D + 2\widetilde{A}_k D\left(2\widetilde{A}_k + \frac{1}{\sqrt{D}}\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{1}{\sqrt{D}}\frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + \sqrt{\widetilde{B}_k}\right) \\
&\leq \left(\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + 2\widetilde{A}_k\sqrt{D} + \sqrt{\widetilde{B}_k D}\right)^2 \\
&= \left(\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + 2A_k + \sqrt{B_k}\right)^2 \qquad (52) \\
&\leq \left(\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + 2A_\infty + \sqrt{B_\infty}\right)^2 \qquad (53)
\end{aligned}$$

if $\left\{\sqrt{\varepsilon_k}\right\}_{k=0}^{\infty}$ is absolutely summable (and therefore, $\{\varepsilon_k\}_{k=0}^{\infty}$ is also absolutely summable), where

$$A_k \triangleq \widetilde{A}_k\sqrt{D} = \sum_{j=1}^{k}\sqrt{\tfrac{\varepsilon_{j-1}}{\left(1-\sigma\tau\|\mathbf{A}\|_2^2\right)\sigma}}\,, \tag{54}$$

and

$$B_k \triangleq \widetilde{B}_k D = \sum_{j=1}^{k}\varepsilon_{j-1}\,. \tag{55}$$

Hence, from (42), we have

$$\frac{\left\|\mathbf{x}^{(k)} - \hat{\mathbf{x}}\right\|_2^2}{2\sigma} + \frac{\left\|\mathbf{z}^{(k)} - \hat{\mathbf{z}}\right\|_2^2}{2\tau} \leq \frac{1}{D}\left(\frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}} + 2A_\infty + \sqrt{B_\infty}\right)^2 < \infty\,. \tag{56}$$

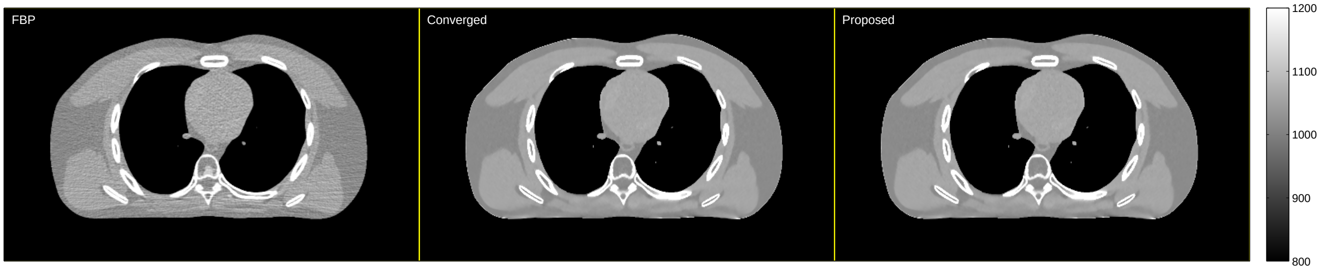

Fig. 1: XCAT phantom: cropped images (displayed from 800 to 1200 HU) from the central transaxial plane of the initial FBP image $\mathbf{x}^{(0)}$ (left), the reference reconstruction $\mathbf{x}^{\star}$ (center), and the reconstructed image using the proposed algorithm (OS-LALM-24-c-1) at the 30th iteration $\mathbf{x}^{(30)}$ (right).

This implies that the sequence of updates $\left\{\left(\mathbf{z}^{(k)}, \mathbf{x}^{(k)}\right)\right\}_{k=0}^{\infty}$ generated by the inexact CP in (27) is a bounded sequence. Let

$$C \triangleq \frac{\left\|\mathbf{x}^{(0)} - \hat{\mathbf{x}}\right\|_2}{\sqrt{2\sigma}} + \frac{\left\|\mathbf{z}^{(0)} - \hat{\mathbf{z}}\right\|_2}{\sqrt{2\tau}}. \tag{57}$$

From (43) and the convexity of h and g^*, we have

$$\begin{aligned}
\Omega\left(\mathbf{z}_k, \hat{\mathbf{x}}\right) - \Omega\left(\hat{\mathbf{z}}, \mathbf{x}_k\right) &\le \frac{1}{k}\sum_{j=1}^{k}\left(\Omega\left(\mathbf{z}^{(j)}, \hat{\mathbf{x}}\right) - \Omega\left(\hat{\mathbf{z}}, \mathbf{x}^{(j)}\right)\right) \\
&\le \frac{\left(C + 2A_k + \sqrt{B_k}\right)^2}{k} \qquad (58) \\
&\le \frac{\left(C + 2A_\infty + \sqrt{B_\infty}\right)^2}{k}, \qquad (59)
\end{aligned}$$

where $\mathbf{z}_k \triangleq \frac{1}{k}\sum_{j=1}^{k}\mathbf{z}^{(j)}$, and $\mathbf{x}_k \triangleq \frac{1}{k}\sum_{j=1}^{k}\mathbf{x}^{(j)}$. That is, the primal-dual gap of $(\mathbf{z}_k, \mathbf{x}_k)$ converges to zero with rate $O(1/k)$. Following the procedure in [2, Section 3.1], we can further show that the sequence of updates $\left\{\left(\mathbf{z}^{(k)}, \mathbf{x}^{(k)}\right)\right\}_{k=0}^{\infty}$ generated by the inexact CP in (27) converges to a saddle-point of (9) if the dimension of $\mathbf{x}$ and $\mathbf{z}$ is finite. ■

II. Additional experimental results

A. XCAT phantom

We simulated an axial CT scan by using a $1024 \times 1024 \times 154$ XCAT phantom [6] for 500 mm transaxial field-of-view (FOV), where $\Delta_x = \Delta_y = 0.4883$ mm and $\Delta_z = 0.6250$ mm. An $888 \times 64 \times 984$ noisy (with Poisson noise) sinogram is numerically generated with GE LightSpeed fan-beam geometry [7] corresponding to a monoenergetic source at 70 keV with 10^5 incident photons per ray and no background event. When reconstructing images, we used a $512 \times 512 \times 90$ image volume with a coarser grid, where $\Delta_x = \Delta_y = 0.9766$ mm and $\Delta_z = 0.6250$ mm, and an edge-preserving regularizer defined in [1] with a scaled Fair potential function $\phi(x) \triangleq \delta^2\left(|t|/\delta - \log(1 + |t|/\delta)\right)$ for $\delta = 10$ HU, a directional regularization parameter β_i that is inversely proportional to the squared distance to the nearest neighbor in each direction, and a voxel-dependent weight $\kappa_n \triangleq \sqrt{[\mathbf{A}'\mathbf{W}\mathbf{1}]_n/[\mathbf{A}'\mathbf{1}]_n}$ [8], where the jth diagonal entry of the diagonal weighting matrix $\mathbf{W}$ is defined as $w_j \triangleq \exp(-y_j)$.

Figure 1 shows the cropped images from the central transaxial plane of the initial FBP image (with Hanning filtering), the reference reconstruction, and the reconstructed image using the proposed algorithm (OS-LALM-24-c-1) at the 30th iteration. Figure 2 and Figure 3 show the reconstructed images and the difference images using the OS+momentum algorithm and the proposed algorithm with different numbers of subsets ($M = 12$, 24, and 36), respectively. As the number of subsets increases, the OS+momentum algorithm becomes less stable and generates noise-like artifacts inside the object, degrading the image quality. In comparison, the proposed algorithm remains stable even with 36 subsets and produces accurate reconstructions. Finally, Figure 4 shows the convergence rate curves of different OS-based algorithms with different numbers of subsets, and we can see that with similar fast convergence rate in first few iterations, the proposed algorithm shows better stability even when using many subsets for acceleration.

B. Shoulder scan with Barzilai-Borwein acceleration

In this experiment, we demonstrated accelerating the proposed algorithm using the Barzilai-Borwein (spectral) method [9] that mimics the Hessian $\mathbf{A}'\mathbf{W}\mathbf{A}$ by $\mathbf{H}_k \triangleq \alpha_k \mathbf{D}_{\mathsf{L}}$. The scaling factor α_k is solved by fitting the secant equation:

$$\mathbf{y}_k \approx \mathbf{H}_k \mathbf{s}_k \tag{60}$$

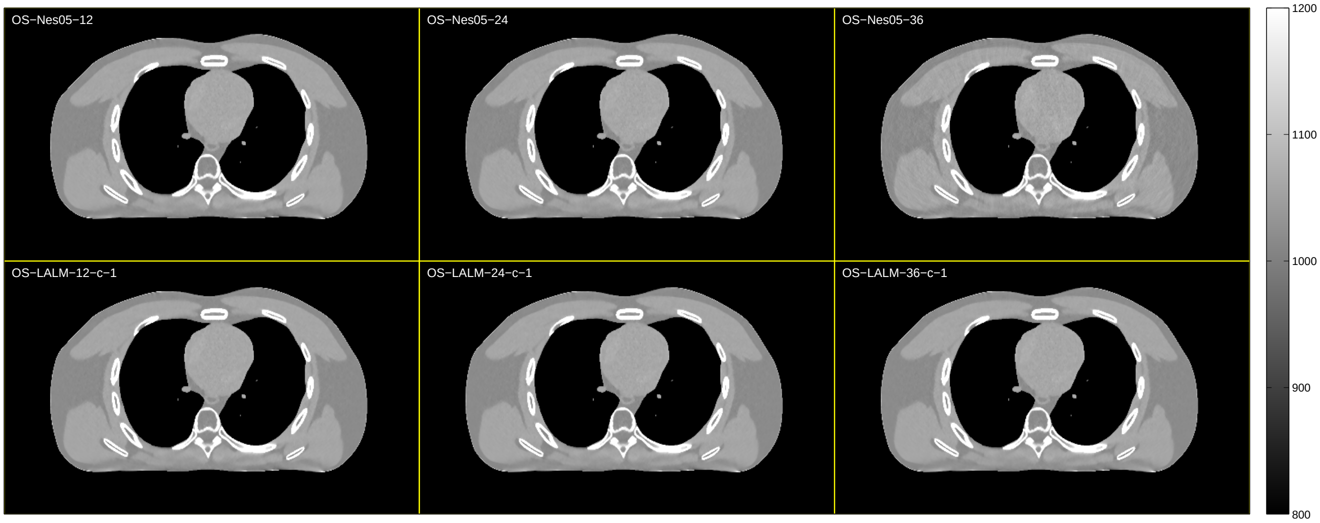

Fig. 2: XCAT phantom: cropped images (displayed from 800 to 1200 HU) from the central transaxial plane of the reconstructed image $\mathbf{x}^{(30)}$ using the OS+momentum algorithm and the proposed algorithm with different numbers of subsets ($M = 12$, 24, and 36).

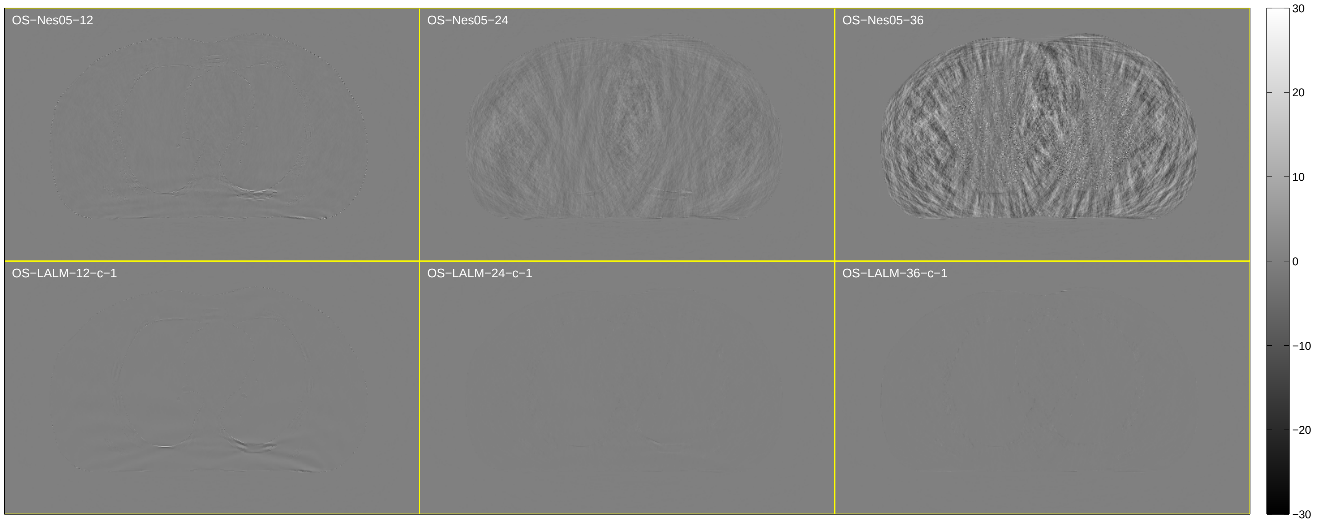

Fig. 3: XCAT phantom: cropped difference images (displayed from -30 to 30 HU) from the central transaxial plane of $\mathbf{x}^{(30)} - \mathbf{x}^{\star}$ using the OS+momentum algorithm and the proposed algorithm with different numbers of subsets ($M = 12$, 24, and 36).

in the weighted least-squares sense, i.e.,

$$\alpha_k = \arg\min_{\alpha \le 1} \tfrac{1}{2} \left\| \mathbf{y}_k - \alpha \mathbf{D}_{\mathsf{L}} \mathbf{s}_k \right\|_{\mathbf{P}}^2 \tag{61}$$

for some positive definite $\mathbf{P}$, where

$$\mathbf{y}_k \triangleq \nabla \mathsf{L}\big(\mathbf{x}^{(k)}\big) - \nabla \mathsf{L}\big(\mathbf{x}^{(k-1)}\big) \tag{62}$$

and

$$\mathbf{s}_k \triangleq \mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\,. \tag{63}$$

We choose $\mathbf{P}$ to be $\mathbf{D}_{\mathsf{L}}^{-1}$ since $\mathbf{D}_{\mathsf{L}}^{-1}$ is proportional to the step sizes of the voxels. By choose $\mathbf{P} = \mathbf{D}_{\mathsf{L}}^{-1}$, we are fitting the secant equation with more weight for voxels with larger step sizes. Note that applying the Barzilai-Borwein acceleration changes $\mathbf{H}_k$ every iteration, and the majorization condition does not necessarily hold. Hence, the convergence theorems developed in Section I are not applicable. However, ordered-subsets (OS) based algorithms typically lack convergence proofs anyway, and

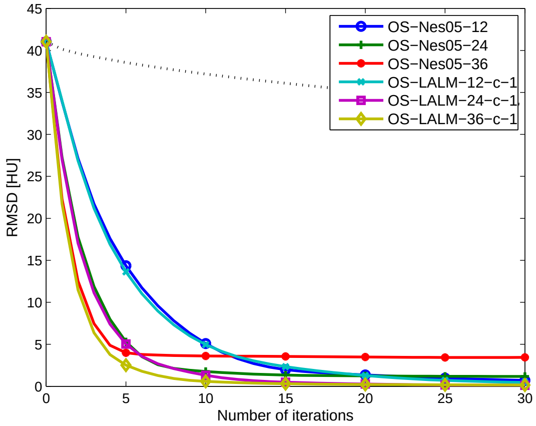

Fig. 4: XCAT phantom: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration using OS-based algorithms with 12, 24, and 36 subsets. The dotted line shows the RMS differences using the standard OS algorithm with one subset.

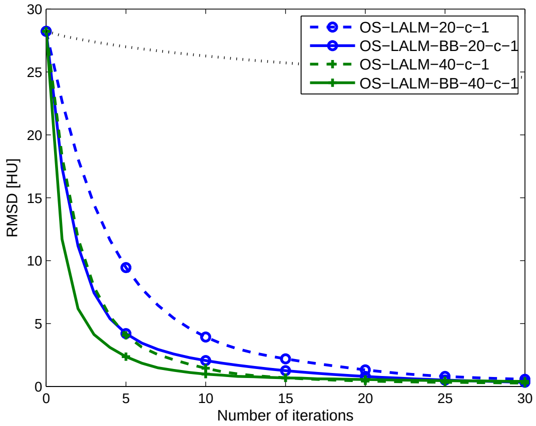

Fig. 5: Shoulder scan: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ as a function of iteration using the proposed algorithm without and with the Barzilai-Borwein acceleration. The dotted line shows the RMS differences using the standard OS algorithm with one subset.

we find that this acceleration works well in practice. Figure 5 shows the RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^{\star}$ of the shoulder scan dataset as a function of iteration using the proposed algorithm without and with the Barzilai-Borwein acceleration. As can be seen in Figure 5, the proposed algorithm with both $M = 20$ and $M = 40$ shows roughly 2-times acceleration in early iterations using the Barzilai-Borwein acceleration.

C. Truncated abdomen scan

In this experiment, we reconstructed a $600 \times 600 \times 239$ image from an abdomen region helical CT scan with transaxial truncation, where the sinogram has size $888 \times 64 \times 3516$ and pitch 1.0. The maximum number of subsets suggested in [1] is about 20. Figure 6 shows the cropped images from the central transaxial plane of the initial FBP image, the reference reconstruction, and the reconstructed image using the proposed algorithm (OS-LALM-20-c-1) at the 30th iteration. This experiment demonstrates how different OS-based algorithms behave when the number of subsets exceeds the suggested maximum number of subsets. Figure 7 shows the difference images for different OS-based algorithms with 10, 20, and 40 subsets. As can be seen in Figure 7, the proposed algorithm works best for $M = 20$; when M is larger ($M = 40$), ripples and light OS artifacts appear. However, it is still much better than the OS+momentum algorithm [10]. In fact, the OS artifacts in the reconstructed image using the OS+momentum algorithm with 40 subsets are visible with the naked eye in the display window from 800 to 1200 HU. The convergence rate curves in Figure 8 support our observation. In sum, the proposed algorithm exhibits fast convergence rate and excellent gradient error tolerance even in the case with truncation.

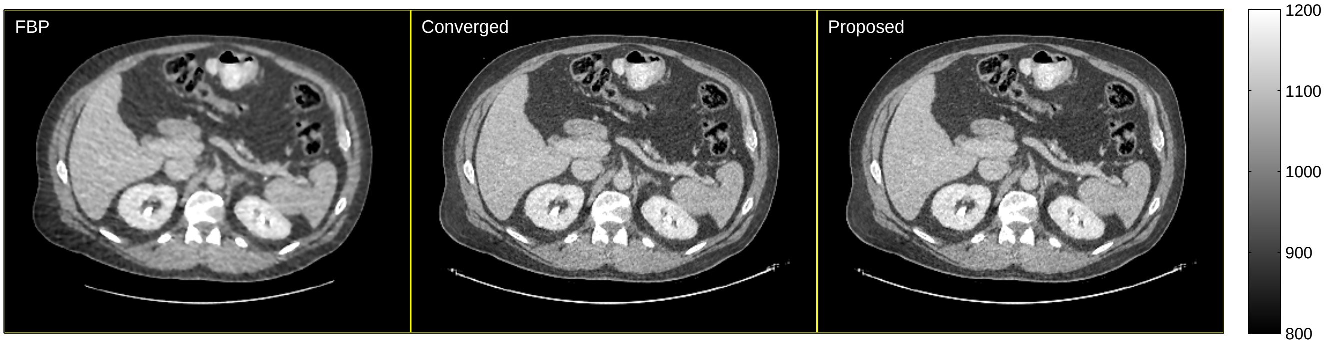

Fig. 6: Truncated abdomen scan: cropped images (displayed from 800 to 1200 HU) from the central transaxial plane of the initial FBP image $\mathbf{x}^{(0)}$ (left), the reference reconstruction $\mathbf{x}^{\star}$ (center), and the reconstructed image using the proposed algorithm (OS-LALM-20-c-1) at the 30th iteration $\mathbf{x}^{(30)}$ (right).

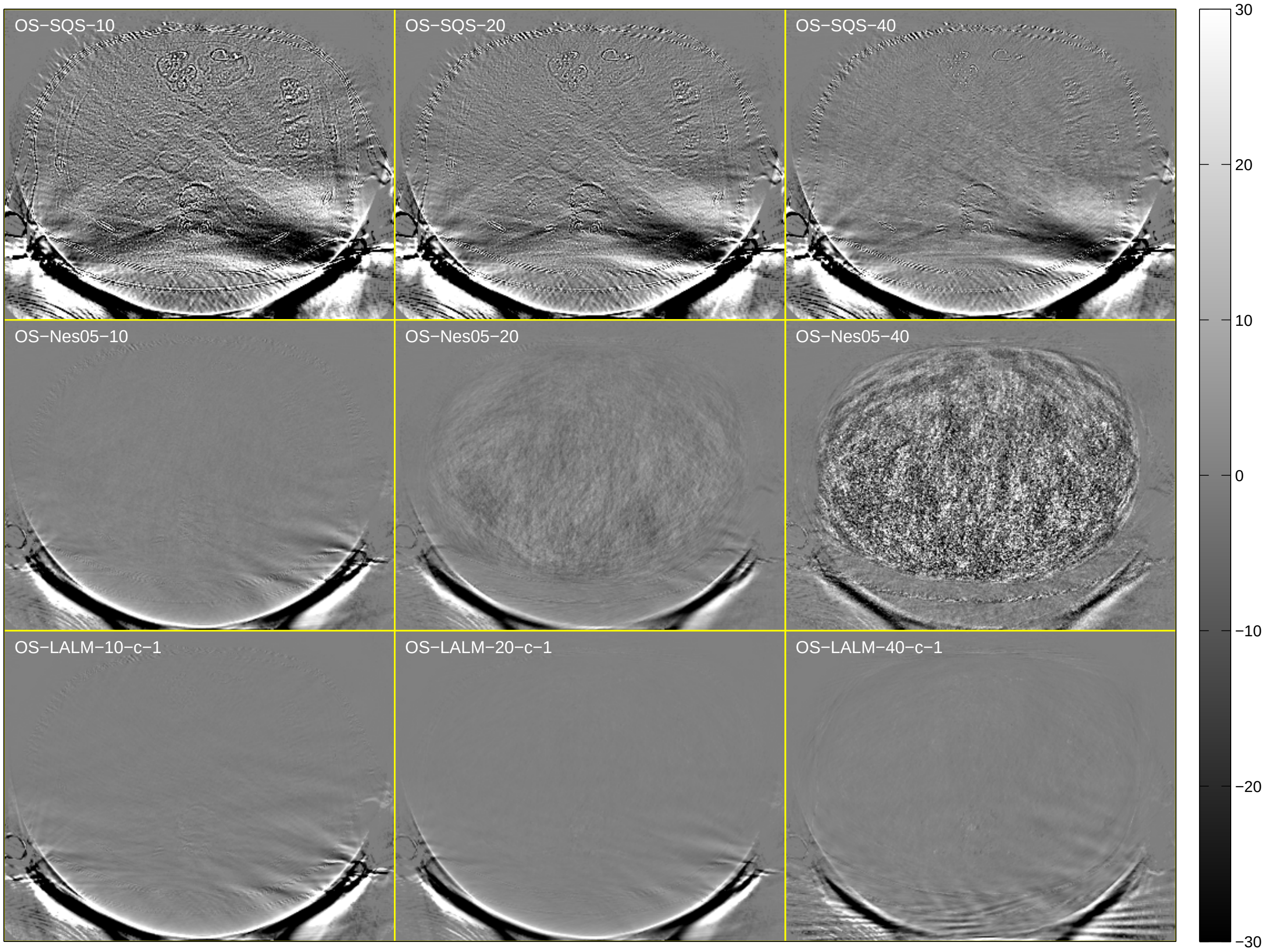

Fig. 7: Truncated abdomen scan: cropped difference images (displayed from −30 to 30 HU) from the central transaxial plane of $\mathbf{x}^{(30)} - \mathbf{x}^{\star}$ using OS-based algorithms.

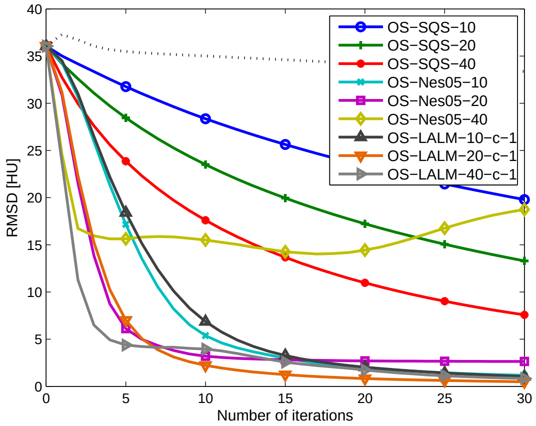

Fig. 8: Truncated abdomen scan: RMS differences between the reconstructed image $\mathbf{x}^{(k)}$ and the reference reconstruction $\mathbf{x}^\star$ as a function of iteration using OS-based algorithms with 10, 20, and 40 subsets. The dotted line shows the RMS differences using the standard OS algorithm with one subset.

Appendix
Outline of the proposed OS-LALM algorithm

We outlined the proposed OS-LALM algorithm for solving PWLS X-ray CT image reconstruction problems:

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\Omega} \left\{ \mathsf{L}(\mathbf{x}) + \mathsf{R}(\mathbf{x}) \right\} , \tag{64}$$

where $\mathsf{L}(\mathbf{x}) \triangleq \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2_{\mathbf{W}}$ is the weighted quadratic data-fitting term, and R is an edge-preserving regularization term. Let $\mathbf{D}_\mathsf{L}$ and $\mathbf{D}_\mathsf{R}$ denote the diagonal majorizing matrices of L and R, respectively, and $[\cdot]_\mathcal{C}$ denote the projection operator onto a convex set $\mathcal{C}$. The proposed OS-LALM algorithm (with downward continuation) is described in Algorithm 1. The inner loop is the fast iterative shrinkage/thresholding algorithm (FISTA) [11] for solving the constrained weighted denoising problem:

$$\hat{\mathbf{z}} \in \arg\min_{\mathbf{z}\in\Omega} \left\{ \frac{1}{2} \left\| \mathbf{z} - \left(\mathbf{x} - (\rho \mathbf{D}_\mathsf{L})^{-1} \mathbf{s}^+\right) \right\|^2_{\rho\mathbf{D}_\mathsf{L}} + \mathsf{R}(\mathbf{z}) \right\} . \tag{65}$$

When $n = 1$, i.e., with a single gradient descent for (65), Algorithm 1 can be further simplified as Algorithm 2 that takes $1/M$ forward/back-projection and one regularizer gradient evaluation per iteration (looping over subsets). The computational complexity of Algorithm 2 is almost the same as standard OS algorithm [12] with negligible overhead.

Algorithm 1: The proposed algorithm (OS-LALM-M-c-n) for solving PWLS X-ray CT image reconstruction problems.

```
Input: $M \geq 1$, $n \geq 1$, and initilize $\mathbf{x}$ by an FBP image.
initialize $\rho = 1$, $m = 1$, $\boldsymbol{\zeta} = \mathbf{g} = M \nabla \mathsf{L}_1(\mathbf{x})$
for $i = 1, 2, \ldots$ do
    $\mathbf{s}^+ = \rho \boldsymbol{\zeta} + (1 - \rho)\, \mathbf{g}$
    initialize $\mathbf{z} = \mathbf{v} = \mathbf{x}$, $\tau = 1$
    for $j = 1, 2, \ldots, n$ do
        $\boldsymbol{\sigma}^+ = \rho \mathbf{D}_\mathsf{L} \left(\mathbf{v} - \left(\mathbf{x} - (\rho \mathbf{D}_\mathsf{L})^{-1} \mathbf{s}^+\right)\right)$
        $\mathbf{z}^+ = \left[\mathbf{v} - (\rho \mathbf{D}_\mathsf{L} + \mathbf{D}_\mathsf{R})^{-1} (\boldsymbol{\sigma}^+ + \nabla \mathsf{R}(\mathbf{v}))\right]_\Omega$
        $\tau^+ = \left(1 + \sqrt{1 + 4\tau^2}\right) / 2$
        $\mathbf{v}^+ = \mathbf{z}^+ + \frac{\tau - 1}{\tau^+} (\mathbf{z}^+ - \mathbf{z})$
    end
    $\mathbf{x}^+ = \mathbf{z}^+$
    $\boldsymbol{\zeta}^+ = M \nabla \mathsf{L}_{m^+ = m+1}(\mathbf{x}^+)$
    $\mathbf{g}^+ = \frac{\rho}{\rho + 1} \boldsymbol{\zeta}^+ + \frac{1}{\rho + 1} \mathbf{g}$
    $\rho^+ = \frac{\pi}{i + 1} \sqrt{1 - \left(\frac{\pi}{2i + 2}\right)^2}$
end
```

Algorithm 2: The proposed algorithm (OS-LALM-M-c-1) for solving PWLS X-ray CT image reconstruction problems.

```
Input: $M \geq 1$ and initilize $\mathbf{x}$ by an FBP image.
initialize $\rho = 1$, $m = 1$, $\boldsymbol{\zeta} = \mathbf{g} = M \nabla \mathsf{L}_1(\mathbf{x})$
for $i = 1, 2, \ldots$ do
    $\mathbf{s}^+ = \rho \boldsymbol{\zeta} + (1 - \rho)\, \mathbf{g}$
    $\mathbf{x}^+ = \left[\mathbf{x} - (\rho \mathbf{D}_{\mathsf{L}} + \mathbf{D}_{\mathsf{R}})^{-1} (\mathbf{s}^+ + \nabla \mathsf{R}(\mathbf{x}))\right]_{\Omega}$
    $\boldsymbol{\zeta}^+ = M \nabla \mathsf{L}_{m^+ = m+1}(\mathbf{x}^+)$
    $\mathbf{g}^+ = \frac{\rho}{\rho+1} \boldsymbol{\zeta}^+ + \frac{1}{\rho+1} \mathbf{g}$
    $\rho^+ = \frac{\pi}{i+1} \sqrt{1 - \left(\frac{\pi}{2i+2}\right)^2}$
end
```

References

[1] H. Nien and J. A. Fessler, “Fast X-ray CT image reconstruction using a linearized augmented Lagrangian method with ordered subsets,” *IEEE Trans. Med. Imag.*, 2014. Submitted.

[2] A. Chambolle and T. Pock, “A first-order primal-dual algorithm for convex problems with applications to imaging,” *J. Math. Im. Vision*, vol. 40, no. 1, pp. 120–45, 2011.

[3] M. V. Afonso, J. M. Bioucas-Dias, and M. A. T. Figueiredo, “An augmented Lagrangian approach to the constrained optimization formulation of imaging inverse problems,” *IEEE Trans. Im. Proc.*, vol. 20, pp. 681–95, Mar. 2011.

[4] J. Eckstein and D. P. Bertsekas, “On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators,” *Mathematical Programming*, vol. 55, pp. 293–318, Apr. 1992.

[5] M. Schmidt, N. Le Roux, and F. Bach, “Convergence rates of inexact proximal-gradient methods for convex optimization,” in *Adv. in Neural Info. Proc. Sys.*, pp. 1458–66, 2011.

[6] W. P. Segars, M. Mahesh, T. J. Beck, E. C. Frey, and B. M. W. Tsui, “Realistic CT simulation using the 4D XCAT phantom,” *Med. Phys.*, vol. 35, pp. 3800–8, Aug. 2008.

[7] J. Wang, T. Li, H. Lu, and Z. Liang, “Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose X-ray computed tomography,” *IEEE Trans. Med. Imag.*, vol. 25, pp. 1272–83, Oct. 2006.

[8] J. A. Fessler and W. L. Rogers, “Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs,” *IEEE Trans. Im. Proc.*, vol. 5, pp. 1346–58, Sept. 1996.

[9] J. Barzilai and J. Borwein, “Two-point step size gradient methods,” *IMA J. Numerical Analysis*, vol. 8, no. 1, pp. 141–8, 1988.

[10] D. Kim, S. Ramani, and J. A. Fessler, “Accelerating X-ray CT ordered subsets image reconstruction with Nesterov’s first-order methods,” in *Proc. Intl. Mtg. on Fully 3D Image Recon. in Rad. and Nuc. Med*, pp. 22–5, 2013.

[11] A. Beck and M. Teboulle, “A fast iterative shrinkage-thresholding algorithm for linear inverse problems,” *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[12] H. Erdoğan and J. A. Fessler, “Ordered subsets algorithms for transmission tomography,” *Phys. Med. Biol.*, vol. 44, pp. 2835–51, Nov. 1999.