Ordered Subsets Acceleration using Relaxed Momentum for X-ray CT Image Reconstruction

Donghwan Kim and Jeffrey A. Fessler

***Abstract*—Statistical image reconstruction in X-ray CT can provide decent images even with low dose, but requires substantial computation time. Recently, we have proposed combining ordered subsets (OS) methods and Nesterov's momentum technique for accelerated X-ray CT image reconstruction. We have observed rapid convergence speed of the proposed algorithms in our experiments, but sometimes encountered unstable behavior. Therefore, we introduce a diminishing step size rule, called a *relaxed* momentum approach, to stabilize the algorithm, while preserving the fast convergence rate. We use a real 3D CT scan to show that the proposed approach can achieve both fast convergence rate and stability.**

I. Introduction

We reconstruct a (nonnegative) image $x \in \mathbb{R}_+^{N_p}$ from a noisy measured sinogram data $y \in \mathbb{R}^{N_d}$ by minimizing a convex and continuously differentiable objective function $\Psi(x)$, based on the statistics of X-ray CT. This paper focuses on a penalized weighted least squares (PWLS) cost function [1]:

$$\hat{x} = \operatorname*{arg\,min}_{x \succeq 0} \left\{ \Psi(x) \triangleq \frac{1}{2}||y - Ax||_W^2 + R(x) \right\}, \qquad (1)$$

where A is a projection operator [2], a diagonal matrix W provides statistical weighting [3], and $R(x)$ is a (nonquadratic and edge-preserving) regularization function.

Iterative algorithms require long computation time for minimizing the 3D CT cost function $\Psi(x)$ in (1). Previously in [4], we applied Nesterov's momentum method [5] to ordered subsets based on separable quadratic surrogates (OS-SQS) algorithms [6], [7] for accelerated convergence rate (without increasing the computational cost per iteration). However, we observed some undesirable instability of the accelerated OS algorithm in some cases. Thus, in this paper, we investigate the diminishing step size rule suggested in [8] to suppress the accumulation of error coming from OS methods.

The convergence analysis of accelerated stochastic gradient methods with momentum is proposed in [8], where a diminishing step size rule provides stabilized convergence even with momentum. In this paper, we treat OS-SQS methods as *diagonally preconditioned* stochastic gradient methods, and propose to adapt the convergence analysis and the diminishing step size rule in [8], which we call a *relaxed* momentum approach. We investigate various schemes for relaxing momentum to achieve overall fast convergence rate (with stability). We use a real 3D phantom scan to verify the stabilizing behavior of the proposed method.

II. Stochastic OS-SQS algorithm

In 3D CT, both forward and back projections A and A' become a computational bottleneck. So, we usually prefer OS algorithms [9] that use only a subset of a measurement data to reduce the computation per image update. In other words, OS methods define the subset gradient:

$$\nabla \Psi_m(x) \triangleq A_m' W_m (A_m x - y_m) + \frac{1}{M} \nabla R(x) \qquad (2)$$

for $m = 0, \cdots, M-1$ where M is the number of subsets, and y_m, A_m, and W_m are sub-matrices of y, A and W. Then, OS methods use the subset gradient $M\nabla\Psi_m(x)$ (with a scaling constant M) instead of $\nabla\Psi(x)$ to reduce computational cost. This enables approximately M times accelerations in run time for early iterations when the following holds

$$\nabla \Psi(x) \approx M \nabla \Psi_m(x). \qquad (3)$$

Larger M would be preferable for faster initial convergence, but OS methods will reach larger limit-cycle that loops around the optimum [10], due to the increased discrepancy between $\nabla\Psi(x)$ and $M\nabla\Psi_m(x)$.

We can view OS methods in a stochastic sense by defining $M\nabla\Psi_{S_k}(x)$ as a stochastic estimate of $\nabla\Psi(x)$, where a random variable S_k at kth iteration is uniformly chosen from $\{0, 1, \cdots, M-1\}$. This stochastic OS algorithm combined with SQS method [6], [7] is illustrated in **Algorithm 1**, where D is a diagonal majorization matrix that satisfies

$$\Psi(x) \le \Psi(\bar{x}) + \nabla\Psi(\bar{x})'(x - \bar{x}) + \frac{1}{2}||x - \bar{x}||_D^2 \qquad (4)$$

for all $x, \bar{x} \in \mathbb{R}_+^{N_p}$. The matrix D can be computed using a Lipschitz constant or SQS methods [6], [7]. The notation $[\cdot]_+$ enforces the nonnegativity constraint by clipping the negative values to zero.

The stochastic estimate $M\nabla\Psi_{S_k}(x)$ of the exact gradient $\nabla\Psi(x)$ in **Algorithm 1** satisfies:

$$E_{S_k}[M\nabla\Psi_{S_k}(x)] = \nabla\Psi(x) \qquad (5)$$

$$E_{S_k}[|M\nabla_j\Psi_{S_k}(x) - \nabla_j\Psi(x)|^2] \le \sigma_j^2, \ \forall\, j, \qquad (6)$$

for $k = 0, 1, 2, \cdots$ and $x \in \mathcal{B}$ where $\mathcal{B}$ is a bounded feasible set[1], $\nabla_j \triangleq \partial/\partial x_j$, and we define a matrix $\Sigma \triangleq \mathsf{diag}\{\sigma_j\}$. The property (6) for a *diagonally-preconditioned* stochastic

D. Kim and J. A. Fessler are with the Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109 USA (e-mail:kimdongh@umich.edu, fessler@umich.edu).

Supported in part by NIH grant R01-HL-098686 and equipment donations from Intel. Helical CT data provided by GE Healthcare.

[1] The property (6) holds if x is in a bounded set. We derive a bounded feasible set $\mathcal{B}$ including the optimum $\hat{x}$ from the measurement data y based on [10, Section A.2], and implicitly enforce the generated sequences of the algorithms to be within the set.

978-1-4799-0534-8/13/$31.00 ©2013 IEEE

Algorithm 1. Stochastic OS-SQS algorithm.

```
Initialize $x^{(0)}$ and compute $D$.
for $n = 0, 1, 2, \cdots$
for $m = 0, 1, \cdots, M-1$
    $k = nM + m$
    Compute $\nabla\Psi_{\xi_k}(x^{(\frac{k}{M})})$, where $\xi_k$ is a realization of $S_k$.
    $x^{(\frac{k+1}{M})} = \left[x^{(\frac{k}{M})} - D^{-1}M\nabla\Psi_{\xi_k}(x^{(\frac{k}{M})})\right]_+$
end
end
```

OS-SQS algorithm is a slightly generalized version of the property of stochastic gradient in [8]. We use these properties for the convergence analysis of the proposed algorithms in next section. Even though it is impractical to estimate the value of σ_j, it is obvious that the elements of matrix $\Sigma = \mathsf{diag}\{\sigma_j\}$ become small by appropriately grouping the subsets and using small M.

Recently, we have combined OS-SQS algorithm and Nesterov's momentum approach [5], called OS-SQS-momentum, significantly accelerating X-ray CT image reconstruction [4]. However, we experienced the accumulation of stochastic error Σ of OS algorithm that leads to unstable convergence behavior. Thus, we adapt a diminishing step size rule developed for stochastic gradient method with momentum for our OS-SQS-momentum algorithm. In next section, we introduce and extend the result in [8] for minimizing X-ray CT cost function (1) rapidly and efficiently.

III. Stochastic OS-SQS algorithm with relaxed momentum

A. Algorithm

Algorithm 2. Stochastic OS-SQS algorithm with momentum.

```
Define $\Gamma^{(k)} \triangleq \mathsf{diag}\left\{\gamma_j^{(k)}\right\}$, $\alpha_0 = 1$, and $\alpha_k \triangleq \max_j \frac{\gamma_j^{(k)}}{\gamma_j^{(k-1)}}$.
Initialize $x^{(0)} = v^{(0)} = z^{(0)}$, $t_0 = 1$, and compute $D$.
for $n = 0, 1, 2, \cdots$
for $m = 0, 1, \cdots, M-1$
    $k = nM + m$
    Compute $\nabla\Psi_{\xi_k}(z^{(\frac{k}{M})})$, where $\xi_k$ is a realization of $S_k$.
    Choose $\Gamma^{(k)}$ s.t. $\begin{cases}\Gamma^{(0)} \succ D, & k = 0\\ \Gamma^{(k)} \succeq \Gamma^{(k-1)}, & k > 0\end{cases}$.
    $t_{k+1} = \frac{1}{2\alpha_k}\left(1 + \sqrt{1 + 4t_k^2\alpha_k\alpha_{k+1}}\right)$
    $x^{(\frac{k+1}{M})} = \left[z^{(\frac{k}{M})} - [\Gamma^{(k)}]^{-1}M\nabla\Psi_{\xi_k}(z^{(\frac{k}{M})})\right]_+$
    $v^{(\frac{k+1}{M})} = \left[z^{(0)} - [\Gamma^{(k)}]^{-1}\sum_{l=0}^{k} t_l M\nabla\Psi_{\xi_l}(z^{(\frac{l}{M})})\right]_+$
    $z^{(\frac{k+1}{M})} = \left(1 - \frac{t_{k+1}}{\sum_{l=\bullet}^{k+1} t_l}\right)x^{(\frac{k+1}{M})} + \frac{t_{k+1}}{\sum_{l=\bullet}^{k+1} t_l}v^{(\frac{k+1}{M})}$
end
end
```

Algorithm 2 illustrates a generalized version of OS-SQS-momentum methods, where the algorithm reduces to the previously proposed version [4, Fig. 4] when we use a deterministic subset ordering $S_k = (k \bmod M)$ and a fixed diagonal matrix

$$\Gamma^{(k)} = D. \tag{7}$$

For $M = 1$, the algorithm with these "conventional" choices can be proven to satisfy

$$\Psi(x^{(\frac{k+1}{M})}) - \Psi(\hat{x}) \leq \frac{2||x^{(0)} - \hat{x}||_D^2}{(k+1)(k+2)} \tag{8}$$

by generalizing the derivation in [5]. By applying OS algorithm with $M > 1$, we were able to achieve M^2 acceleration in early iterations [4], by replacing k with $nM+m$ considering the computational cost in OS algorithm. Note that the computation cost of **Algorithm 2** is similar to that of **Algorithm 1** even though it looks more complicated, because it uses one full projection and back-projection per outer iteration (n) as **Algorithm 1**.

However, for $M > 1$, the analysis in [8] (with a stochastic variable S_k) illustrates that the choice (7) might suffer from the accumulation of error from OS methods as:

$$E[\Psi(x^{(\frac{k+1}{M})}) - \Psi(\hat{x})] \leq \frac{2||x^{(0)} - \hat{x}||_D^2}{(k+1)(k+2)} + \frac{(k+3)\,\mathsf{tr}\{P\Sigma\}}{3}, \tag{9}$$

where the matrix $P \triangleq \mathsf{diag}\left\{p_j \triangleq \max_{x,\bar{x}\in\mathcal{B}} |x_j - \bar{x}_j|\right\}$ measures the diameter of the feasible set $\mathcal{B}$. The error Σ coming from OS method affects the last term in (9), which we want to suppress to improve stability.

Using the larger fixed $\Gamma^{(k)} = qD$ with $q > 1$ did not help prevent the accumulation of error [8]. Therefore, to better stabilize the algorithm, here we adapt the *relaxed* momentum approach in [8] that increases the denominator in **Algorithm 2** as follows:

$$\Gamma^{(k)} = D + (k+2)^c\,\Gamma \tag{10}$$

for any choice of a constant $c \geq 0$ and a diagonal matrix $\Gamma \succ 0$. With the properties (5) and (6), we can achieve the following inequality (11) by extending the result in [8]:

Lemma *1:* For a constant $c \in [0, 2]$, the sequence $\{x^{(\frac{k+1}{M})}\}$ generated by **Algorithm 2** with (10) satisfies

$$E[\Psi(x^{(\frac{k+1}{M})}) - \Psi(\hat{x})] \leq \left(\max_{0\leq l\leq k}\sqrt{\alpha_l}\right)\left[\frac{2||x^{(0)} - \hat{x}||_D^2}{(k+1)(k+2)} + \frac{2||x^{(0)} - \hat{x}||_\Gamma^2}{(k+1)(k+2)^{1-c}} + \frac{2(k+3)^{3-c}\,\mathsf{tr}\{\Gamma^{-1}\Sigma^2\}}{(3-c)(k+1)(k+2)}\right]. \tag{11}$$

Proof: See the proof in [11]. ■

The coefficient c controls the rate of accumulating error in (11), and we investigate the choices of c in next section.

B. The choice of c

Among the possible choices of $c \in [0,\ 2]$, we consider two cases $c = 1$ and $c = 1.5$ for better understanding of the convergence analysis in (11).

We first consider the choice $c = 1$ that provides the rate

$$O\left(\frac{2||x^{(0)} - \hat{x}||_D^2}{k^2} + \frac{2||x^{(0)} - \hat{x}||_\Gamma^2}{k} + \frac{\mathsf{tr}\{\Gamma^{-1}\Sigma^2\}}{2}\right) \quad (12)$$

on average. This choice prevents the accumulation of error but does not guarantee convergence on average. We can understand from (11) that the coefficient c should be larger than 1 to decrease the accumulated error.

We also consider the choice $c = 1.5$ that provides a rate

$$O\left(\frac{2||x^{(0)} - \hat{x}||_D^2}{k^2} + \frac{2||x^{(0)} - \hat{x}||_\Gamma^2}{\sqrt{k}} + \frac{2\,\mathsf{tr}\{\Gamma^{-1}\Sigma^2\}}{1.5\sqrt{k}}\right) \quad (13)$$

and converges on average, but somewhat slower than (12) in early iterations considering the larger constant $\max_{0\le l\le k}\sqrt{\alpha_l}$ in (11) for same Γ. The choice $c = 1.5$ provides the optimal rate of decrease $O(1/\sqrt{k})$ of the stochastic noise for the first-order methods [8], [12]. The parameter c that is larger than 1.5 seems not useful as it does not have the optimal convergence rate.

In the result section, we consider the choices $c = 1$ and $c = 1.5$ providing the rates (12) and (13). An optimal choice of matrix Γ providing fast convergence rate based on (11) remains unknown in practice, so we provide a practical approach to choose a reasonable Γ in the result section.

Overall, the proposed *relaxed* momentum approach will eventually reach smaller cost function value than the previous (unrelaxed) choice $\Gamma^{(k)} = D$ (or $\Gamma = 0$) in (7), since we prevent the accumulation of error from OS methods by increasing the denominator $\Gamma^{(k)}$ as (10). In other words, the algorithm with $M > 1$ and the relaxation (10) may be slower than the choice of (7) initially, but eventually becomes faster and reaches closer to the optimum on average.

IV. Results

We used a helical cone-beam CT scan of the GE performance phantom (GEPP) to examine the performance of proposed OS algorithm with relaxed momentum. We reconstructed a $512\times512\times47$ image of GEPP from a $888\times64\times3071$ noisy sinogram data measured in a helical geometry with pitch 0.5.

We used $M = 24$ for OS methods. The convergence analysis in Section III was for stochastic subset ordering in OS methods, but we used the deterministic order of subsets suggested in [13] here, which is known to be a good choice in tomography problems.[2] We leave the more detailed discussion of stochastic OS methods for the upcoming work in [11].

The matrix Γ in (10) controls the constant of the convergence rate in (11). Here, we use the following

$$\Gamma = \gamma D \quad (14)$$

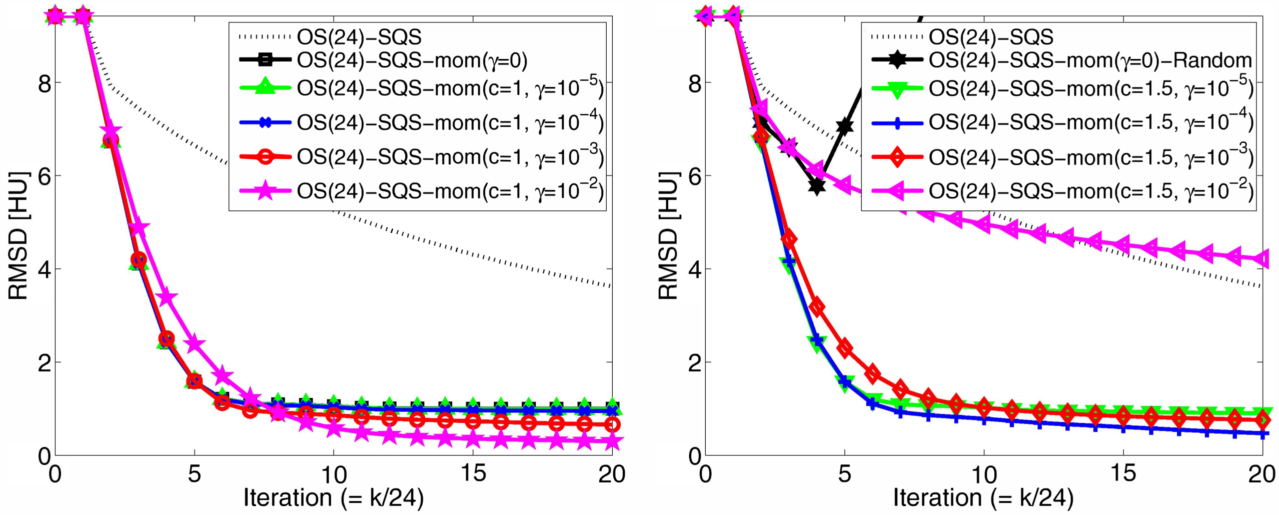
for $\gamma > 0$. (The matrix D generated by [6] for this experiment has maximum value 6.3×10^{12} and median 1.1×10^{11}.) The matrix Γ can be better optimized if we know Σ and the distance between $x^{(0)}$ and $\hat{x}$ for each voxel based on (11), which we will further discuss in [11]. Here, we simply tuned the parameter γ in (14) within $\{0, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\}$.

Fig. 1 shows the root mean square difference (RMSD) between the current and converged image within the region-of-interest (ROI) in Hounsfield Units (HU), versus iteration:

$$\text{RMSD}(x^{(n)}) = ||x_{\text{ROI}}^{(n)} - \hat{x}_{\text{ROI}}||/\sqrt{N_{p,\text{ROI}}}, \quad (15)$$

where $N_{p,\text{ROI}}$ is the number of voxels within the ROI. We count one iteration per 24 sub-iterations (m) for $M = 24$

[2] We found that the subset ordering in [13] greatly prevents the accumulation of error from OS in the proposed algorithm, compared to other subset orderings. For the case shown in Fig. 1, the unrelaxed version ($\gamma = 0$) of OS-SQS-momentum using a random subset ordering becomes highly unstable in Fig. 1(b) as discussed in (9), whereas the algorithm using the subset ordering in [13] behaved relatively stable as seen in Fig. 1(a).



(a) c=1

(b) c=1.5

Fig. 1. Plots of RMSD versus iteration. Computing the matrix D in [6], [7] requires one each forward and back projection and this resulted in no updates at the first iteration.

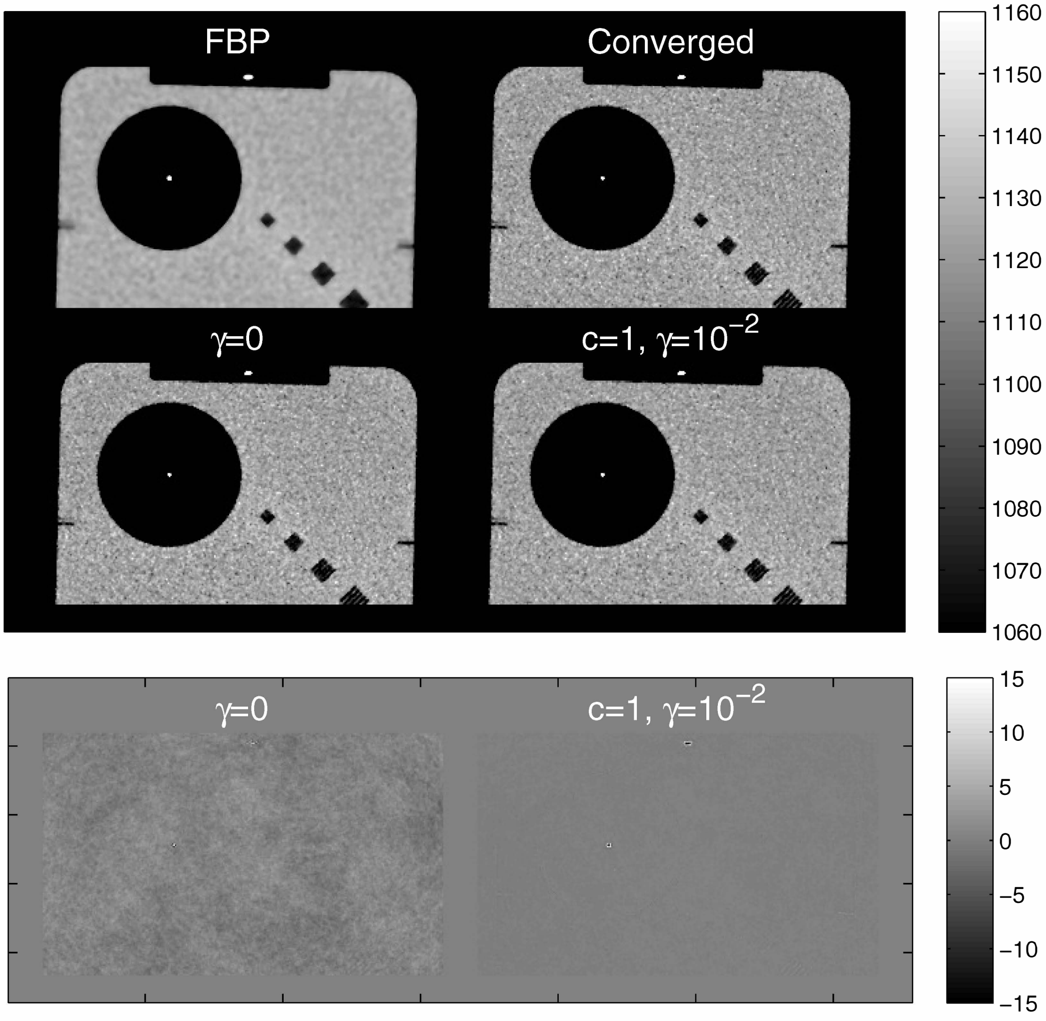


Fig. 2. Top: center slice of FBP image $x^{(0)}$, converged image $\hat{x}$, and reconstructed images at 20th iteration. Bottom: center slice of difference images between the reconstructed and converged image. (Images are cropped for better visualization.)

by considering the computational cost in OS algorithm. The plots illustrate that the all momentum approaches accelerate the OS-SQS algorithm. We can further verify that the relaxed momentum approach becomes more stable and reaches nearby the optimum with a slight sacrifice on the initial convergence rate, compared to the unrelaxed version ($\gamma = 0$). The choice $c = 1$ and $\gamma = 10^{-2}$ provided the fastest overall convergence rate among other choices in this experiment. However, parameter tuning of c and Γ needs to be further (automatically) optimized to achieve *robust* fast convergence rate, which we further discuss in [11].

In Fig. 1, the choice $c = 1.5$ performed worse than the choice $c = 1$ even though we expected the case $c = 1.5$ to achieve the optimal asymptotic rate of decreasing the stochastic noise component of the upper bound on the cost function decrease in (11). As mentioned before, this was due to the larger constant $\max_{0 \le l \le k} \sqrt{\alpha_l}$ for larger c in (11) that slowed down the initial convergence. But, we observed that the choice $c = 1.5$ is in decreasing phase even after 20 iterations while $c = 1$ reaches nondecreasing phase after 15 iterations as discussed in (12). So, this gives a room for further investigation on using $c = 1$ initially and switching to $c = 1.5$ after first few iterations, which we leave as a future work.

Fig. 2 presents the initial filtered back-projection (FBP) image $x^{(0)}$, the converged image $\hat{x}$, and reconstructed images at 20th iteration. We can observe the improvement using the relaxed momentum by looking at the difference images in Fig. 2, where the proposed method reduced the noise texture. However, the relaxed momentum method had not yet finished updating some structures as seen in Fig. 2. This raises the need of a voxel-dependent Γ using $|x_j^{(0)} - \hat{x}_j|$ and σ_j for each jth voxel based on (11). These are unavailable in practice, but some heuristic methods such as an idea in [7] might be useful, which we will investigate in [11].

V. Conclusion

In this paper, we attempted to stabilize the OS-momentum algorithm that were previously found to have fast convergence rate but unstable in some cases. We adapted the relaxation scheme of momentum approach to prevent the accumulation of error while preserving the rapid convergence speed, which we verified on a real 3D CT scan.

The tuning parameter c and Γ that we used in the result section was suboptimal, and we plan to investigate better choices in near future.

REFERENCES

[1] 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, no. 11, pp. 4526–44, Nov. 2007.

[2] Y. Long, J. A. Fessler, and J. M. Balter, “3D forward and back-projection for X-ray CT using separable footprints,” *IEEE Trans. Med. Imag.*, vol. 29, no. 11, pp. 1839–50, Nov. 2010.

[3] K. Sauer and C. Bouman, “A local update strategy for iterative reconstruction from projections,” *IEEE Trans. Sig. Proc.*, vol. 41, no. 2, pp. 534–48, Feb. 1993.

[4] 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*, 2013, pp. 22–5.

[5] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Mathematical Programming*, vol. 103, no. 1, pp. 127–52, May 2005.

[6] H. Erdoğan and J. A. Fessler, “Ordered subsets algorithms for transmission tomography,” *Phys. Med. Biol.*, vol. 44, no. 11, pp. 2835–51, Nov. 1999.

[7] 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, no. 11, pp. 1965–78, Nov. 2013.

[8] O. Devolder, “Stochastic first order methods in smooth convex optimization,” 2011.

[9] H. M. Hudson and R. S. Larkin, “Accelerated image reconstruction using ordered subsets of projection data,” *IEEE Trans. Med. Imag.*, vol. 13, no. 4, pp. 601–9, Dec. 1994.

[10] S. Ahn and J. A. Fessler, “Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms,” *IEEE Trans. Med. Imag.*, vol. 22, no. 5, pp. 613–26, May 2003.

[11] D. Kim, S. Ramani, and J. A. Fessler, “Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction,” In preparation.

[12] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization*, John Wiley, 1983.

[13] G. T. Herman and L. B. Meyer, “Algebraic reconstruction techniques can be made computationally efficient,” *IEEE Trans. Med. Imag.*, vol. 12, no. 3, pp. 600–9, Sept. 1993.