A fast weighted stochastic gradient descent algorithm for image reconstruction in 3D computed tomography

Davood Karimi, Rabab Ward
Department of Electrical and Computer Engineering
University of British Columbia

Nancy Ford
Faculty of Dentistry
University of British Columbia

***Abstract*—We describe and evaluate an algorithm for image reconstruction in 3D x-ray computed tomography. The proposed algorithm is similar to the class of projected gradient methods. The gradient descent for reducing the measurement misfit term is carried out using a stochastic gradient iteration and the gradient directions are weighted using weights suggested by parallel coordinate descent. In addition, to further improve the speed of the algorithm, at each iteration we minimize the cost function on a small subspace spanned by the direction of the current projected gradient and several previous update directions. We apply the proposed algorithm on simulated and real cone-beam projections and compare it with Fast Iterative Shrinkage-Thresholding Algorithm (FISTA).**

I. Introduction

Statistical and iterative reconstruction methods for x-ray computed tomography (CT) have received renewed interest in recent years. The majority of iterative reconstruction algorithms proposed for 3D CT in recent years are based on the class of projected gradient methods. Each iteration of these algorithms involves a gradient step to reduce the measurement misfit term followed by a proximal operator for the regularization term, which is usually the total variation. Accelerated versions of these algorithms, such as FISTA, Nesterov's method, and ADMM have been successfully deployed for 3D CT [1, 2, 3].

In this study, we propose a new algorithm that is different from the basic gradient projection scheme in several ways. For reducing the measurement misfit term, we suggest a parallel coordinate descent update that will lead to a weighted gradient descent step. After applying the proximal operator for the TV regularization, this can be used as the new estimate of the image. However, in the spirit of methods such as the method of conjugate gradients, we use this direction along with the directions of several previous image updates to define a subspace over which the cost function is approximately minimized in each iteration. We apply the proposed algorithm on simulated and real CT data and compare it with FISTA.

II. Materials and Methods

We denote the unknown image by $x \in \mathbb{R}^n$ and the projection measurements by $y \in \mathbb{R}^m$. Our goal is to recover an estimate of the unknown image by solving the following unconstrained problem:

$$\hat{x} = \arg\min_x \frac{1}{2}||Ax - y||_2^2 + \lambda \text{TV}(x) \tag{1}$$

A common approach to solving such a problem is an iterative algorithm that applies, in each iteration, a gradient descent step for the measurement consistency term followed by a proximity operator for the non-smooth regularization term. This basic approach can be extended and accelerated in many ways [4]. Here, we follow a slightly different approach by applying a parallel coordinate descent for the measurement consistency term. Assuming that x^k is our current estimate, if we want to minimize this term with respect to its ith coordinate, the exact solution will be:

$$x^{k+1}[i] = x^k[i] + \frac{A_i^T(y - Ax^k)}{||A_i||^2} \tag{2}$$

where A_i denotes the ith column of A. While a straightforward implementation of this algorithm is possible in theory, it will be impractical for 3D CT because of the very large size of the problem and because fast forward and back-projection algorithms process all measurements in one projection view at once. Therefore, we suggest the following parallel coordinate descent iteration:

$$x^{k+1} = \Psi_\lambda \left(x^k + W^{-1}A^T(y - Ax^k)\right) \tag{3}$$

where $W \in \mathbb{R}^{n \times n}$ is a diagonal matrix whose diagonal elements are norms of the columns of A and Ψ_λ is the proximal operator of TV, i.e.:

$$\Psi_\lambda(u) = \arg\min_x \left(\frac{1}{2}||x - u||_2^2 + \lambda \text{TV}(x)\right) \tag{4}$$

The iteration in (3) is in the form of forward-backward splitting algorithms. A similar iteration was suggested

in [5] for ℓ_1-regularized problems. However, these algorithms are known to have a slow convergence rate. To improve their speed, several algorithms have been proposed in recent years. In most of these algorithms (e.g. [1, 3]), the speedup is achieved by exploiting the directions of previous updates. In other words, x^{k+1} is computed not only based on x^k, but also x^{k-1}. In sequential subspace optimization, proposed in [6], directions of several previous updates are exploited and superior convergence rates are reported. This idea is very similar in essence to the method of conjugate gradients applied to quadratic functions. Following this idea, at every iteration of the algorithm we minimize the cost function in (1) over the subspace spanned by the direction suggested by (3) as well as the directions of K previous updates. Formally:

$$x^{k+1} = x^k + \sum_{i=1}^{K+1} \hat{\alpha}_i^k d_i^k$$

where

$$\begin{cases} d_1^k = x^k - \Psi_\lambda \left(x^k + W^{-1} A^T (y - Ax^k) \right) \\ d_i^k = x^{k-i+1} - x^{k-i} \quad i = 2 : K+1 \end{cases} \tag{5}$$

and the coefficients α_is are chosen to minimize the cost function:

$$\hat{\alpha}_i^k = \arg\min_{\alpha} \frac{1}{2} \left|\left| A(x^k + \sum_{i=1}^{K+1} \alpha_i^k d_i^k) - y \right|\right|_2^2 + \lambda \mathrm{TV}\left(x^k + \sum_{i=1}^{K+1} \alpha_i^k d_i^k \right)$$

This minimization is not easy because TV is non-smooth. The approach followed in this study was to sequentially minimize the cost function with respect to individual α_is using golden section search. In our experience, the values of α_is did not change drastically between successive iterations of the algorithm. Therefore, α_is can be initialized to the values found in the previous iteration and a search performed in a small neighborhood around these values. This way, a single sweep through α_is, starting with α_1, was enough to find $\hat{\alpha}_i$s to good accuracy.

The inner update in (3) (i.e. the update prior to the application of the proximal operator, which also appears or in computing d_1^k in (5)) is in the form of a weighted gradient descent. The algorithm can be made significantly faster by replacing this step with a stochastic gradient descent algorithm [7, 8]. To explain this idea, let us note that the measurement misfit term can be written as:

$$F(x) = \frac{1}{2} ||Ax - y||_2^2 = \sum_{i=1}^{n} ||A_i x - y_i||_2^2 = \sum_{i=1}^{n} f_i(x)$$

where n is total number of projection views, y_i is the measurements in the ith projection, and A_i is the sub-matrix of A contaning only those rows that correspond to the ith projection. This form of a cost function is very conducive to stochastic gradient descent. A full gradient descent iteration for this cost function will have this form:

$$x^{k+1} = x^k + \gamma_k A^T (y - Ax^k) \tag{7}$$

which is the same as the inner update in (3) save the multiplication by the diagonal matrix W, whereas a stochastic gradient descent step will be:

$$x^{k+1} = x^k + \gamma_k A_{i_k}^T (y_{i_k} - Ax_{i_k}^k) \tag{8}$$

where i_k is a randomly selected index from among the set $\{1, ..., n\}$ and γ_k is the step size.

Therefore, at each iteration of the algorithm, instead of performing a full gradient step, we perform n stochastic gradient descents, where N is the number of projection views. The order of projection views is chosen randomly in each iteration and each projection view is used exactly once in each iteration. Stochastic gradient descents usually exhibit fast convergence in the initial iterations but their convergence rate deteriorates as the algorithm makes progress [9]. This is because the direction of a stochastic gradient is equal to the direction of the true (full) gradient only in expectation and the variance can be quite high. Therefore, it is common to use diminishing step sizes. We use a rule of the form $\gamma_k = \gamma_0/(1+\gamma_0 \beta k)$, where γ_0 is the initial step size, k is an iteration number, and β is a decay parameter [9]. We used a value of $\beta = 10$ which we found empirically. As for the initial value, γ_0, used different values for different projection views. Specifically, γ_{0_i} is selected to be inversely proportional to largest eigenvalue of the corresponding sub-matrix A_i, which is the Lipschitz constant of the gradient of the measurement misfit term associated with a projection view (i.e., $f_i(x)$). These eigenvalues can be estimated using a power method and stored before the start of the algorithm.

In order to evaluate the proposed algorithm, we applied it on a set of simulated data and two sets of real cone-beam projections. The simulated data consisted of 90 projections with uniform angular spacing between 0° and 178° from a 3D Shepp-Logan phantom. A phantom of size $256 \times 256 \times 256$ voxels and a flat detector of 360×360 pixels were considered. The incident photon count was considered to be $N_0 = 2 \times 10^4$. The real cone-beam sinograms were acquired using a Gamma Medica eXplore CT 120 micro-CT scanner. The imaged objects included a phantom, designed in [10] for comprehensive evaluation of the performance of micro-CT scanners, and a dead rat. The flat panel detector

included 3500×2272 detector elements, which with 4×4 binning generated 875×568 sinograms. The scan of the phantom consisted of 110 projections at $1.754°$ intervals, whereas the scan of the rat consisted of 122 projections at $1.484°$ intervals. The tube voltage, tube current, and exposure times were, respectively, 80 kV, 32 mA, and 8 ms for the scan of the phantom, and 50 kV, 63 mA, and 100 ms for the rat scan. The size of the reconstructed image was $880 \times 880 \times 650$ voxels, with isotropic voxels of $0.1 \times 0.1 \times 0.1$ mm^3 in size. We compared the proposed algorithm with Fast Iterative Shrinkage- Thresholding Algorithm (FISTA) [1]. Every iteration of the proposed method and FISTA require one forward-projection and one back-projection, which account for the main computational load. Therefore, we use the iteration count as a measure of computational effort. An important parameter in the proposed algorithm is the number of previous update directions, K. In all the experiments reported in this study we used $K = 3$.

III. Results

For the simulated data from the Shepp-Logan phantom, Figure 1 shows the Root-mean-square of the reconstruction error (RMSE), where reconstruction error is defined as the difference between the reconstructed and true images, for the proposed algorithm and FISTA for up to 45 iterations. Both algorithms were initialized with a filtered-backprojection reconstruction. The proposed algorithm has a better start but the two algorithms get closer with more iterations.

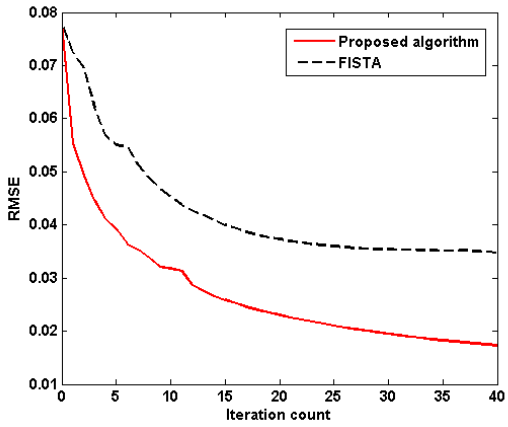

Figure 1. Change in the RMS of the reconstruction error for the Shepp-Logan phantom reconstructed with the proposed algorithm and FISTA.

The physical phantom imaged using the micro-CT scanner contained a module including a slanted edge that consisted of a plastic-air boundary. This provided an ideal means of estimating the modulation transfer function, MTF. The estimated MTF for the images reconstructed using the proposed algorithm and FISTA are shown in Figure 2. In Figure 3 we have shown cross sections of the reconstructed phantom at the location of a set of resolution coils and one-dimensional profiles through the coils. We also used a uniform polycarbonate plate in the phantom to assess the noise level in the reconstructed images. The signal-to-noise-ratio for the images reconstructed using FBP, FISTA, and the proposed algorithm were 19.6, 22.3, and 22.7 dB, respectively. Overall, for the construction of the image of the physical phantom, the proposed algorithm performs slightly better than FISTA. As expected, both FISTA and the proposed algorithm outperform FBP.

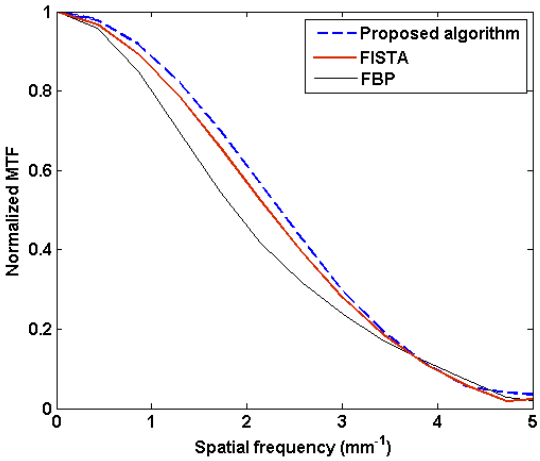

Figure 2. Estimated modulation transfer function for the images of the physical phantom reconstructed using different algorithms.

Figure 4 shows two selected regions of the images of the rat reconstructed with different algorithms. Again, the proposed algorithm achieves a slightly better reconstruction than FISTA.

IV. Discussion

The proposed algorithm shows a promising performance on simulated and real data. A limitation of the proposed algorithm is the need to store several previous update directions and their projections. Each update direction has the size of the reconstructed image and the size of its projection is equal to the size of the projection measurements used in the reconstruction. Considering the large size of 3D CT images and their projections, the memory requirements may be problematic. The required memory grows linearly with the number, K, of previous update directions used. In [5] up to $K = 7$ are used, but in our experience no additional improvements are achieved beyond $K = 3$ and the memory requirements for storing three update directions and their projections should not be prohibitive for most applications.

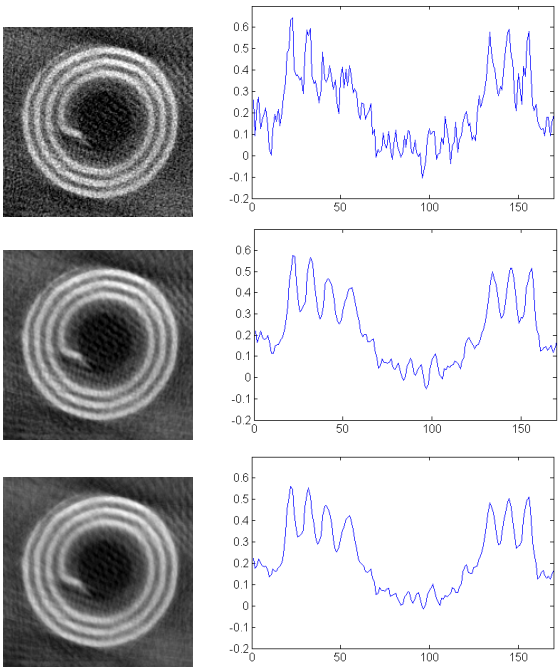

Figure 3. Two-dimensional and one-dimensional profiles of a resolution coil in the physical phantom reconstructed with different algorithms: Top: filtered back-projection, center: FISTA, bottom: proposed algorithm.

V. Acknowledgments

Micro-CT imaging was performed in the Centre for High-Throughput Phenogenomics at the University of British Columbia, a facility supported by the Canada Foundation for Innovation, British Columbia Knowledge Development Foundation, and the UBC Faculty of Dentistry.

References

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

[2] Y. Nesterov, "Gradient methods for minimizing composite objective function," Universite Catholique de Louvain, Tech. Rep. CCIT 559, 2007.

[3] J. Bioucas-Dias and M. A. T. Figueiredo, "A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration," *Image Processing, IEEE Transactions on*, vol. 16, no. 12, pp. 2992–3004, 2007.

[4] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," in *Fixed-point algorithms for inverse problems in science and engineering*. Springer New York, 2011, pp. 185–212.

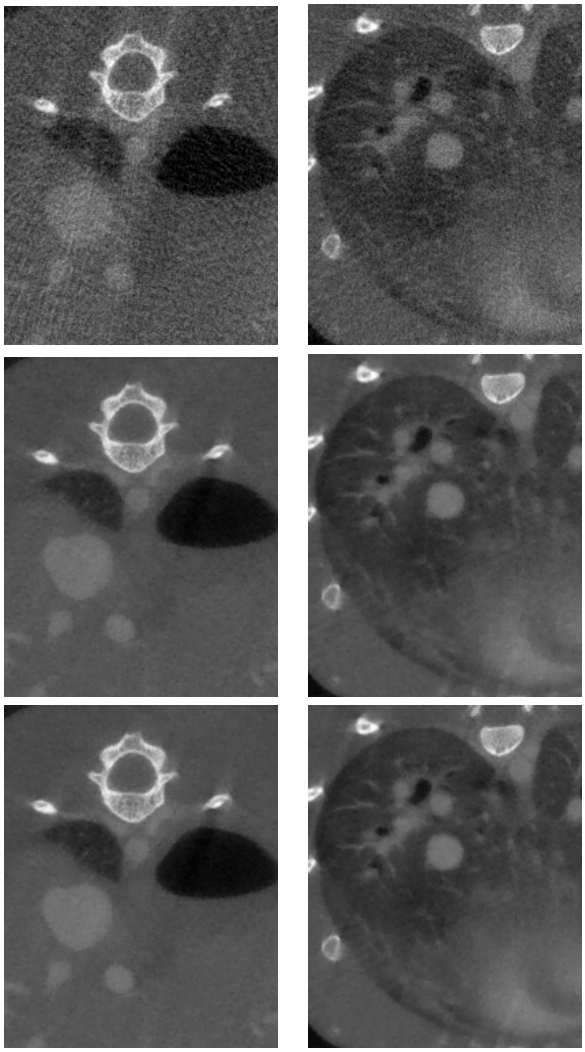

Figure 4. Two-dimensional and one-dimensional profiles of a resolution coil in the physical phantom reconstructed with different algorithms: Top: filtered back-projection, center: proposed algorithm, bottom: FISTA.

[5] M. Zibulevsky and M. Elad, "L1-l2 optimization in signal and image processing," *Signal Processing Magazine, IEEE*, vol. 27, no. 3, pp. 76–88, 2010.

[6] M. Elad, B. Matalon, and M. Zibulevsky, "Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization," *Applied and Computational Harmonic Analysis*, vol. 23, no. 3, pp. 346 – 367, 2007.

[7] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," *arXiv preprint arXiv:1202.6258*, 2012.

[8] A. Defazio, F. Bach, and S. Lacoste-Julien, "Saga: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*,

2014, pp. 1646–1654.

[9] L. Bottou, "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, ser. Lecture Notes in Computer Science, G. Montavon, G. Orr, and K.-R. Mller, Eds. Springer Berlin Heidelberg, 2012, vol. 7700, pp. 421–436.

[10] L. Y. Du, J. Umoh, H. N. Nikolov, S. I. Pollmann, T. Y. Lee, and D. W. Holdsworth, "A quality assurance phantom for the performance evaluation of volumetric micro-CT systems," *Physics in Medicine and Biology*, vol. 52, no. 23, pp. 7087–7108, 2007.