# Nonsmooth nonconvex LDCT image reconstruction via learned descent algorithm

Zhang, Qingchao, Ye, Xiaojing, Chen, Yunmei

Qingchao Zhang, Xiaojing Ye, Yunmei Chen, "Nonsmooth nonconvex LDCT image reconstruction via learned descent algorithm," Proc. SPIE 11840, Developments in X-Ray Tomography XIII, 1184013 (9 September 2021); doi: 10.1117/12.2597798

Event: SPIE Optical Engineering + Applications, 2021, San Diego, California, United States

# Nonsmooth Nonconvex LDCT Image Reconstruction via Learned Descent Algorithm

Qingchao Zhang[a], Xiaojing Ye[b], and Yunmei Chen[a]

[a]Department of Mathematics, University of Florida, Gainesville, USA
[b]Department of Mathematics and Statistics, Georgia State University, Atlanta, USA

## ABSTRACT

Deep neural network architectures based on unrolling optimization algorithms have been widely adopted in deep-learning based image reconstruction applications in recent years. However, these architectures only mimic the iterative schemes of the corresponding algorithms, but lack rigorous convergence guarantee; and the learned network layers are difficult to interpret. These issues have hindered their applications in clinical use. In this paper, we develop an efficient Learned Descent Algorithm with a Line Search strategy (LDA-LS) and apply it to the nonconvex nonsmooth optimization problem of low-dose CT (LDCT) reconstruction. We show that LDA-LS yields a highly interpretable neural network architecture, where the regularization parameterized as multilayer perception is explicitly integrated into the iterative scheme and learned during the training process. We demonstrate that LDA-LS retains convergence guarantee as classical optimization algorithms, while achieving improved efficiency and accuracy in LDCT image reconstruction problems.

**Keywords:** Learned descent algorithm, line search, nonsmooth nonconvex optimization, low-dose CT.

## 1. INTRODUCTION

Computed Tomography (CT) is a medical imaging technology widely used in clinical diagnosis nowadays. CT employs X-ray measurements from different angles to generate cross-sectional images of the human body.[1,2] Low-dose CT (LDCT) is an important means to reduce patient exposure to harmful X-ray but results in various degrees of noise and artifacts which must be removed by adaptive image reconstruction method for clinical use.

Filtered Back-Projection (FBP) is a classic analytical method to reconstruct image from limited projection data, however, the reconstruction quality is severely degraded with substantial noise and artifacts in LDCT scenario. To reconstruct better images from LDCT, a number of methods including preprocessing, postprocessing, and hybrid methods have been proposed in the literature.[3–7]

In recent years, we have witnessed tremendous success of deep neural networks, which is at the heart of deep learning, in a large variety of real-world imaging applications. In particular, convolutional neural networks (CNNs) have been applied to CT image reconstruction on sparse view and low dose data.[6–12] However, deep learning methods are widely criticized for being difficult to interpret and data demanding. A recent strategy known as unrolling method aims at mitigating these issues and gained popularity in image reconstruction applications.[13–17] Unrolling method constructs a deep neural network that mimic the iterative structure of some known optimization scheme, such as proximal gradient descent algorithm, but replaces the proximal operator with a multilayer perceptron which is to be learned during training. Nevertheless, the resemblance between the obtained deep neural network and the original optimization scheme is superficial, and such unrolling method lacks convergence guarantees that are of paramount importance in both theory and practice.

In this paper, we adopt the framework called learnable descent algorithm (LDA),[18] and enhance it with a Line Search Strategy to avoid explicitly using the Lipschitz constant of the objective function which is generally unknown in practice. We then apply LDA-LS to the LDCT image reconstruction problem. The algorithm is termed as LDA-LS. LDA-LS parameterizes the regularization function as the composition of the $l_{2,1}$ norm and a

Q.Z.: E-mail: qingchaozhang@ufl.edu
X.Y.: E-mail: xye@gsu.edu
Y.C.: E-mail: yun@ufl.edu

Developments in X-Ray Tomography XIII, edited by Bert Müller
Ge Wang, Proc. of SPIE Vol. 11840, 1184013 · © 2021 SPIE
CCC code: 0277-786X/21/$21 · doi: 10.1117/12.2597798

smooth but nonconvex feature mapping in the form of a deep CNN. Due to the presence of $l_{2,1}$ norm and CNN, the objective function be minimized is nonsmooth and nonconvex. LDA-LS leverages the Nesterov's smoothing technique and the idea of residual learning to arrive at a descent-type algorithm. This algorithm is provably convergence with explicit iteration complexity analysis.[18] In this work, we provide the details of LDA-LS and demonstrate its promising performance in LDCT image reconstruction.

The remainder of the paper is organized as follows. In section 2, we present the related works in the literature that associate with our problem. Then in section 3, we present our method by first defining our model and each of its components, and then stating the algorithm and details of network training. Section **??** presents the numerical results including parameter study, ablation study and comparison with other competing algorithms.

## 2. RELATED WORKS

Deep learning methods have been successfully applied to CT image reconstruction in the past several years.[8–10, 19–22] In,[20] a Residual Encoder-Decoder Convolutional Neural Network (RED-CNN) is developed to reconstruct low-dose CT images using normal dose training data. In,[8] FBPConvNet is developed based on U-net[23] for LDCT reconstruction. More structured deep neural networks including unrolling methods are proposed.[5, 13, 14, 16, 24–29] For example, in,[14] a deep network architecture called BCD-Net is developed which further improves reconstruction quality and generalizes better than methods such as FBPConvNet. Momentum structure has also been integrated into the network architecture in.[13, 27] To address lack of convergence and further improve performance of unrolling methods, in,[18] a Learned Descent Algorithm (LDA) is developed. The LDA architecture is fully determined by the algorithm and thus the network is fully interpretable. As interpretability and convergence guarantee is highly desirable in medical imaging, this framework is a promising method for inverse problems such as LDCT reconstruction.

## 3. METHOD

In this section, we provide the details of the LDA-LS and associated parameter training strategy for LDCT image reconstruction. We first present the image reconstruction problem with learnable regularization in Section 3.1. The nonsmooth noncvonex regularization used in our model and its smooth approximation are given in Section 3.2. Section 3.3 presents LDA-LS and its convergence result.

### 3.1 Image reconstruction with learnable regularization

Image reconstruction is a typical inverse problem that can be modeled in a variational form and cast as an optimization problem as follows:

$$\min_{\mathbf{x}\in\mathcal{X}} \ \phi(\mathbf{x};\mathbf{b},\boldsymbol{\theta}) := f(\mathbf{x};\mathbf{b}) + r(\mathbf{x};\boldsymbol{\theta}), \tag{1}$$

where $\mathbf{x}$ is the image to be reconstructed, $\mathcal{X}$ is the admissible set of $\mathbf{x}$ (e.g., $\mathcal{X} = \mathbb{R}^n$ and $n$ is the number of pixels in $\mathbf{x}$), $f$ is the data fidelity term of $\mathbf{x}$, and $r$ represents the regularization term to overcome the illposedness of the problem. In LDCT, we choose $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ as the data-fidelity term, where $\mathbf{A}$ is the system matrix for CT scanner and $\mathbf{b}$ is the (noisy) sinogram measurements, and $\|\cdot\|$ is the standard Euclidean norm. It is worth pointing out that our method can be readily extended to general smooth but (possibly) nonconvex $f$. In (1), $\boldsymbol{\theta}$ stands for the learnable parameter that determines the regularization function $r$. In particular, we choose to parameterize $r$ as the composition of the sparsity-promoting $l_{2,1}$ norm and multiple convolutional layers. The convolutional layers form a CNN which plays the role of a sparse feature mapping, and the $l_{2,1}$ norm is employed here since it is very powerful for robust group sparse feature selection. We provide more details on the structure of $r$ in the following subsection. We expect that the parameter $\boldsymbol{\theta}$ learned from training data yields a much more adaptive regularization $r$ than handcrafted ones that are often overly simplified and cannot capture the complex structures of images.

To learn the parameter $\boldsymbol{\theta}$ of the regularization $r$, we leverage a set of training data and form the parameter learning problem as a bi-level optimization:

$$\min_{\boldsymbol{\theta}} \quad \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{s=1}^{N} \|\mathbf{x}^{(s)}(\boldsymbol{\theta}) - \hat{\mathbf{x}}^{(s)}\|^2, \tag{2a}$$

$$\text{s.t.} \quad \mathbf{x}^{(s)}(\boldsymbol{\theta}) \in \operatorname*{arg\,min}_{\mathbf{x}\in\mathcal{X}} \{\phi(\mathbf{x};\mathbf{b}^{(s)},\boldsymbol{\theta}) = f(\mathbf{x};\mathbf{b}^{(s)}) + r(\mathbf{x};\boldsymbol{\theta})\}. \tag{2b}$$

That is, the optimal parameter $\boldsymbol{\theta}$ is obtained by minimizing the loss function $\mathcal{L}$, where $\mathcal{L}$ measures the average squared error between $\mathbf{x}^{(s)}(\boldsymbol{\theta})$, the minimizer of the objective function $\phi(\cdot;\mathbf{b}^{(s)},\boldsymbol{\theta})$ with LDCT data $\mathbf{b}^{(s)}$ using the parameter $\boldsymbol{\theta}$, and the high-quality ground-truth image $\hat{\mathbf{x}}^{(s)}$ corresponding $\mathbf{b}^{(s)}$. Here $N$ is the number of training data pairs $(\mathbf{b}^{(s)}, \hat{\mathbf{x}}^{(s)})$. For notation simplicity, we write $f(\mathbf{x})$ and $r(\mathbf{x})$ instead of $f(\mathbf{x};\mathbf{b}^{(s)})$ and $r(\mathbf{x};\boldsymbol{\theta})$ respectively hereafter.

### 3.2 Parametric form of learnable regularization

In this work, we employ the parametric form of $r$ in (2) as the composition of the group sparsity promoting function $l_{2,1}$ norm and a deep CNN architecture. Specifically, we formulate $r$ as follows:

$$r(\mathbf{x}) = \|\mathbf{g}(\mathbf{x})\|_{2,1} = \sum_{i=1}^{m} \|\mathbf{g}_i(\mathbf{x})\|, \tag{3}$$

where each $\mathbf{g}_i(\mathbf{x}) \in \mathbb{R}^d$ is as a feature vector at the position $i \in [m]$. In our experiments, we set the feature extraction operator $\mathbf{g}$ to the following $l$-layer CNN with nonlinear activation function $\sigma$ as follows:

$$\mathbf{g}(\mathbf{x}) = \mathbf{w}_l * \sigma \cdots \; \sigma(\mathbf{w}_3 * \sigma(\mathbf{w}_2 * \sigma(\mathbf{w}_1 * \mathbf{x}))), \tag{4}$$

where $\{\mathbf{w}_q\}_{q=1}^{l}$ denote the convolution weights each consisting of $d$ kernels with identical spatial kernel size $(3 \times 3)$, and $*$ denotes the convolution operation. Here, the componentwise activation function $\sigma$ is set to a smoothed rectified linear unit (ReLU):

$$\sigma(x) = \begin{cases} 0, & \text{if } x \le -\delta, \\ \frac{1}{4\delta}x^2 + \frac{1}{2}x + \frac{\delta}{4}, & \text{if } -\delta < x < \delta, \\ x, & \text{if } x \ge \delta, \end{cases} \tag{5}$$

where $\delta$ is set to be 0.001 throughout the experiments in this work. The smooth activation function $\sigma$ in (5) grants a smooth but nonconvex feature mapping $\mathbf{g}$. Hence, the gradient $\nabla\mathbf{g}$ can be computed in a straightforward manner, where each $\{\mathbf{w}_q^{\top}\}$ can be implemented as transposed convolutional operators.[30] However, since the $l_{2,1}$ norm in (3) is not differentiable, the optimization problem (1) or (2b) is nonsmooth and nonconvex . To tackle this problem, we apply the Nesterov's smoothing technique[31] to get the smooth approximation:[18]

$$r_\varepsilon(\mathbf{x}) = \sum_{i\in I_0} \frac{1}{2\varepsilon}\|\mathbf{g}_i(\mathbf{x})\|^2 + \sum_{i\in I_1}\left(\|\mathbf{g}_i(\mathbf{x})\| - \frac{\varepsilon}{2}\right), \tag{6}$$

where $I_0 = \{i \in [m] \mid \|\mathbf{g}_i(\mathbf{x})\| \le \varepsilon\}$, $I_1 = [m] \setminus I_0$. Here the parameter $\varepsilon$ controls how close the smoothed $r_\varepsilon(\mathbf{x})$ is to the original function $r(\mathbf{x})$, and one can readily show that $r_\varepsilon(\mathbf{x}) \le r(\mathbf{x}) \le r_\varepsilon(\mathbf{x}) + \frac{m\varepsilon}{2}$ for all $\mathbf{x}$ in $\mathbb{R}^n$. From (6) we can calculate $\nabla r_\varepsilon(\mathbf{x})$:

$$\nabla r_\varepsilon(\mathbf{x}) = \sum_{i\in I_0} \nabla\mathbf{g}_i(\mathbf{x})^{\top}\frac{\mathbf{g}_i(\mathbf{x})}{\varepsilon} + \sum_{i\in I_1}\nabla\mathbf{g}_i(\mathbf{x})^{\top}\frac{\mathbf{g}_i(\mathbf{x})}{\|\mathbf{g}_i(\mathbf{x})\|}, \tag{7}$$

where $\nabla\mathbf{g}_i(\mathbf{x}) \in \mathbb{R}^{d\times n}$ is the Jacobian of $\mathbf{g}_i$ at $\mathbf{x}$. As we will show in the next subsection, the smoothing parameter $\varepsilon$ is automatically reduced and gradually tends to 0 during the iterations, such that $r_\varepsilon$ can closely approximate the original nonsmooth regularization $r$.

### 3.3 Learned descent algorithm

The training of network parameter $\boldsymbol{\theta}$ is outlined as follows. To mitigate the challenges in solving the bi-level optimization problem (2), we approximate the solution $\mathbf{x}^{(s)}(\boldsymbol{\theta})$ of the lower-level problem (2b) by iterating the Learned Descent Algorithm with Line Search (presented in Algorithm 1), or LDA-LS in short, for a fixed number $K$ of iterations. The last iterate, denoted by $\mathbf{x}_K^{(s)}(\boldsymbol{\theta})$, can be thought of as an approximate solution of (2b) and used in the upper-level problem (2a). Moreover, the LDA-LS is applied to the lower-level problem (2b) with the smoothed regularization $r_\varepsilon$ given in (6) since explicit calculation of gradients is required. The optimal parameter $\boldsymbol{\theta}$ is obtained by minimizing the upper-level problem (2a) using a stochastic gradient descent method such as ADAM such that $\mathbf{x}_K^{(s)}(\boldsymbol{\theta})$ is close to the ground truth image in the training data set.

We now provide the details of the derivation of LDA-LS and its parameter training algorithm. We first focus on the lower-level problem (2b) which LDA-LS is intended to solve. In each iteration $k$, we apply a proximal gradient descent step to (2b) with a smoothed regularization $r_\varepsilon$ in (6) and $\varepsilon = \varepsilon_k$ as follows:

$$\mathbf{z}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k), \tag{8a}$$

$$\mathbf{x}_{k+1} = \operatorname{prox}_{\alpha_k r_{\varepsilon_k}}(\mathbf{z}_{k+1}), \tag{8b}$$

where the proximal operator is defined as

$$\operatorname{prox}_{\alpha r}(\mathbf{z}) := \arg\min_{\mathbf{x}} \left\{ \frac{1}{2\alpha}\|\mathbf{x}-\mathbf{z}\|^2 + r(\mathbf{x}) \right\}.$$

However, the proximal operator does not have a closed-form solution due to the complex structure of $r_{\varepsilon_k}$. Therefore, we approximte $r_{\varepsilon_k}$ by

$$\tilde{r}_{\varepsilon_k}(\mathbf{x}) = r_{\varepsilon_k}(\mathbf{z}_{k+1}) + \langle \nabla r_{\varepsilon_k}(\mathbf{z}_{k+1}), \mathbf{x}-\mathbf{z}_{k+1}\rangle + \frac{1}{2\beta_k}\|\mathbf{x}-\mathbf{z}_{k+1}\|^2 \tag{9}$$

with some $\beta_k > 0$ and obtain a closed-form approximation of (8b) as follows:

$$\mathbf{u}_{k+1} = \operatorname{prox}_{\alpha_k \tilde{r}_{\varepsilon_k}}(\mathbf{z}_{k+1}) = \mathbf{z}_{k+1} - \tau_k \nabla r_{\varepsilon_k}(\mathbf{z}_{k+1}), \tag{10}$$

where $\tau_k = \frac{\alpha_k\beta_k}{\alpha_k+\beta_k}$. We also compute a safeguard iterate $\mathbf{v}_{k+1}$ as

$$\mathbf{v}_{k+1} = \arg\min_{\mathbf{x}} \langle \nabla f(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \langle \nabla r_{\varepsilon_k}(\mathbf{x}_k), \mathbf{x}-\mathbf{x}_k\rangle + \frac{1}{2\alpha_k}\|\mathbf{x}-\mathbf{x}_k\|^2, \tag{11}$$

which also has a closed-form

$$\mathbf{v}_{k+1} = \mathbf{x}_k - \alpha_k\nabla f(\mathbf{x}_k) - \alpha_k \nabla r_{\varepsilon_k}(\mathbf{x}_k). \tag{12}$$

To ensure objective function decay, we employ a line search strategy by shrinking $\alpha_k$ by $\rho \in (0,1)$ until the following condition holds:

$$\phi_{\varepsilon_k}(\mathbf{v}_{k+1}) - \phi_{\varepsilon_k}(\mathbf{x}_k) \le -\tau\|\mathbf{v}_{k+1}-\mathbf{x}_k\|^2. \tag{13}$$

That is, we check whether $\mathbf{v}_{k+1}$ satisfies (13): if yes, we move on to the next step (14) below; otherwise, we set $\alpha_k$ to $\rho\alpha_k$ and recompute $\mathbf{v}_{k+1}$ using (12). It is straightforward to verify that this line search terminates within finitely many steps due to the Lipschitiz continuity of $\nabla\phi_{\varepsilon_k}$.

Finally, we choose between $\mathbf{u}_{k+1}$ and $\mathbf{v}_{k+1}$ that has the smaller function value $\phi_{\varepsilon_k}$ to be the next iterate $\mathbf{x}_{k+1}$:

$$\mathbf{x}_{k+1} = \begin{cases} \mathbf{u}_{k+1} & \text{if } \phi_{\varepsilon_k}(\mathbf{u}_{k+1}) \le \phi_{\varepsilon_k}(\mathbf{v}_{k+1}), \\ \mathbf{v}_{k+1} & \text{otherwise.} \end{cases} \tag{14}$$

The LDA with Line Search derived above is summarized in Algorithm 1 (LDA-LS). Line 9 of Algorithm 1 presents a *reduction criterion*. That is, if the reduction criterion $\|\nabla\phi_{\varepsilon_k}(\mathbf{x}_{k+1})\| < \sigma\gamma\varepsilon_k$ is satisfied, then the smoothing parameter $\varepsilon_k$ is shrunk by $\gamma \in (0,1)$. In our implementation, the parameter $\boldsymbol{\theta}$ includes all the step sizes $\alpha_k$ and

**Algorithm 1** Learnable Descent Algorithm with Line Search (LDA-LS) for (1)

```
1:  **Input:** Initial $\mathbf{x}_0$, $\rho, \gamma \in (0,1)$, and $\varepsilon_0, \sigma, \tau > 0$. Maximum iteration $K$ or tolerance $\epsilon_{\text{tol}} > 0$.
2:  **for** $k = 0, 1, 2, \ldots, K$ **do**
3:      $\mathbf{z}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$
4:      $\mathbf{u}_{k+1} = \mathbf{z}_{k+1} - \tau_k \nabla r_{\varepsilon_k}(\mathbf{z}_{k+1})$
5:      **repeat**
6:          $\mathbf{v}_{k+1} = \mathbf{x}_k - \alpha_k \nabla \phi_{\varepsilon_k}(\mathbf{x}_k)$ and set $\alpha_k \leftarrow \rho \alpha_k$
7:      **until** $\phi_{\varepsilon_k}(\mathbf{v}_{k+1}) - \phi_{\varepsilon_k}(\mathbf{x}_k) \le -\tau \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2$
8:      If $\phi(\mathbf{u}_{k+1}) \le \phi(\mathbf{v}_{k+1})$, then set $\mathbf{x}_{k+1} = \mathbf{u}_{k+1}$, otherwise set $\mathbf{x}_{k+1} = \mathbf{v}_{k+1}$.
9:      If $\|\nabla \phi_{\varepsilon_k}(\mathbf{x}_{k+1})\| < \sigma\gamma\varepsilon_k$, set $\varepsilon_{k+1} = \gamma\varepsilon_k$; otherwise, set $\varepsilon_{k+1} = \varepsilon_k$.
10:     If $\sigma\varepsilon_k < \epsilon_{\text{tol}}$, terminate.
11: **end for**
12: **Output:** $\mathbf{x}_{k+1}$.
```

$\tau_k$ and the initial smoothing parameter $\varepsilon_0$. Given $N$ training data pairs $\{(\mathbf{b}^{(s)}, \hat{\mathbf{x}}^{(s)})\}_{s=1}^N$ of the ground truth data $\hat{\mathbf{x}}^{(s)}$ and its corresponding LDCT measurement $\mathbf{b}^{(s)}$, the optimal $\boldsymbol{\theta}$ is obtained by minimizing the objective function (2a) with each $\mathbf{x}^{(s)}$ approximated by the $K$th iterate $\mathbf{x}_K^{(s)}$ of LDA-LS as described above.

One of the major advantages of Algorithm 1 over existing unrolling based methods is that LDA-LS has guaranteed convergence with explicit iteration complexity bound similar to.[18] To ensure the convergence, LDA-LS requires a few mild assumptions on $f$ and $\mathbf{g}$ as follows: (A1) $f$ is differentiable and (possibly) nonconvex, and $\nabla f$ is $L_f$-Lipschitz continuous. (A2) Every component of $\mathbf{g}$ is differentiable and (possibly) nonconvex, $\nabla \mathbf{g}$ is $L_g$-Lipschitz continuous, and $\sup_{\mathbf{x}\in\mathcal{X}} \|\nabla \mathbf{g}(\mathbf{x})\| \le M$ for some constant $M > 0$. (A3) $\phi$ is coercive, and $\phi^* = \min_{\mathbf{x}\in\mathcal{X}} \phi(\mathbf{x}) > -\infty$. It is easy to verify that these assumptions readily hold in most image reconstruction problems.[18] Due to the nonsmooth and nonconvex nature of (2b), we need to characterize its optimality condition using Clarke stationary point.[32] Specifically, the Clarke subdifferential and Clarke stationary point are defined as follows

Definition 3.1 (Clarke subdifferential). *Suppose that $f : \mathbb{R}^n \to (-\infty, +\infty]$ is locally Lipschitz, the Clarke subdifferential $\partial f(\mathbf{x})$ of $f$ at $\mathbf{x}$ is defined as*

$$\partial f(\mathbf{x}) := \left\{ \mathbf{w} \in \mathbb{R}^n \mid \langle \mathbf{w}, \mathbf{v} \rangle \le \limsup_{\mathbf{z}\to\mathbf{x},\, t\downarrow 0} \frac{f(\mathbf{z} + t\mathbf{v}) - f(\mathbf{z})}{t}, \forall\, \mathbf{v} \in \mathbb{R}^n \right\}$$

Definition 3.2 (Clarke stationary point). *For a locally Lipschitz function $f$, a point $\mathbf{x} \in R^n$ is called a Clarke stationary point of $f$ if $0 \in \partial f(\mathbf{x})$.*

The following theorem states the convergence of Algorithm 1 (LDA-LS) to a Clarke stationary point of (1). The proof closely follows[18] and hence is omitted here. It is worth noting that the difference from[18] is that Algorithm 1 eliminates the explicit requirement on step size $\alpha_k$ in[18] which can be difficult to estimate in practice. Instead, Algorithm 1 (LDA-LS) employs line search of $\alpha_k$ and thus the descent criterion is automatically satisfied in practice and the convergence can be guaranteed in theory.

Theorem 3.3. *Suppose that $\{\mathbf{x}_k\}$ is the sequence generated by Algorithm 1 with any initial $\mathbf{x}_0$, then the algorithm terminates within $O(\epsilon_{\text{tol}}^{-3})$ steps. If $\epsilon_{\text{tol}} = 0$ and $K = \infty$, and let $\{\mathbf{x}_{k_l+1}\}$ be the subsequence where the reduction criterion Line 7 of Algorithm 1 is met for $k = k_l$ and $l = 1, 2, \ldots,$ then $\{\mathbf{x}_{k_l+1}\}$ has at least one accumulation point, and every accumulation point of $\{\mathbf{x}_{k_l+1}\}$ is a Clarke stationary point of (1).*

## 4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

Each phase (block) of the network forward propagation can be viewed as one algorithm iteration, which motivates us to imitate the iterating of the optimization algorithm and use a stair training strategy.[18] We minimize the loss for 200 epochs each stair using the Adam Optimizer[33] with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and initial learning rate $10^{-4}$. The Xavier method[34] is called to initialize the convolution weights and the smoothing parameter $\varepsilon_0$ is initialized to be 0.001. The input $\mathbf{x}_0$ is obtained by FBP.[35] The default architecture configuration is set as follows: the feature extraction operator $\mathbf{g}$ consists of 4 convolutions with 48 kernels and kernel size $3 \times 3$, then with batch size 2 for training and default phase number 15. The algorithm is implemented using the PyTorch toolbox.[36]

The experiments are performed on two well-known public datasets: the Low Dose CT Image and Projection Data of The Cancer Imaging Archive (TCIA) Public Access[37] and another open source dataset available at the National Biomedical Imaging Archive (NBIA). We randomly selected 10 patients out of 151 patients in TCIA data. Among these 10 patients, we randomly sampled 400 scans from 8 patients for training and another 100 scans from the rest 2 patients for testing. To validate the generalizability, we sampled another testing data from NBIA which consists of 80 images of different parts of the human body for diversity. All sampled images were uniformly resized to $256 \times 256$ and normalized to $[0, 1]$. The distance-driven algorithm[38,39] is adopted to obtain the projection of fan beam CT. In the simulation configuration, we set both source and detector to rotation center distances 25 cm and physical region $17\times17$ cm$^2$. We uniformly cast 1024 projection views in 360° range. The X-ray was detected by 512 detector elements of width 0.72 mm each. To make the simulation alike the real clinical condition, the noisy transmission measurement $I$ was produced by adding Poisson and electronic noise[40]

$$I = Possion(I_0 \exp{(-\hat{b})}) + Normal(0, \sigma_e^2), \tag{15}$$

where $I_0$ is the number of incident photons, $\hat{b}$ is the noise-free projection and $\sigma_e^2$ represents the variance the background electronic noise. Full dose intensity $I_0$ is default to be $1.0\times10^6$[41] and for the equipment measurement error variance $\sigma_e^2$ is constantly 10 for all different dose cases. We obtained the noisy projection $\mathbf{b}$ by taking the logarithm transformation over $I_0/I$. Among low dose cases we simulated three projections of levels 10%, 5% and 2.5% with incident intensity $I_0 = 1.0 \times 10^5, 5.0 \times 10^4$ and $2.5 \times 10^4$ accordingly. All the experiments were done on a server with AMD Ryzen Threadripper 1900X CPU, 32 GB of memory and Nvidia RTX-2080Ti GPUs.

## 4.1 Ablation study

### 4.1.1 Comparison with standard gradient descent

We investigate the effectiveness of the proposed LDA-LS algorithm by comparing with unrolling the standard gradient descent (GD) iteration of (2), and an accelerated inertial version by setting $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \nabla\phi(\mathbf{x}_k) + \theta_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ where $\theta_k$ is also learned (AGD). The result is shown in Figure 1. The PSNR score of LDA-LS at each iteration is much higher than standard GD and AGD, where the latter two have comparable performance.

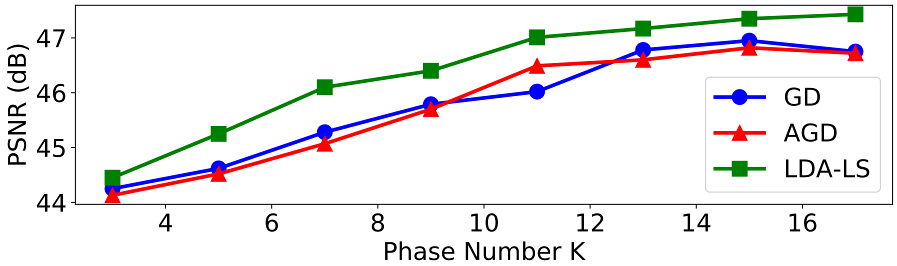


Figure 1. The reconstruction PSNR plots across various GD-type algorithms on TCIA dataset with dose level 10%.

### 4.1.2 Hyper-parameter selection

We check the influence of some hyper-parameters of the architecture by perturbing one while the others remain default if not explicitly mentioned, including the number of convolutions ($l$), the depth of the convolution kernels ($d$) and the phase number ($K$). The default configuration is described at the beginning of Section 4. The impacts of these parameters to the testing results are shown in Figure 1 and Table 1 with dose level 10%. We can observe that the default configuration achieves a good balance between the reconstruction performance and network complexity as desired.

Table 1. The results with different depth of convolution kernels and number of convolutions on TCIA data when $K = 7$.

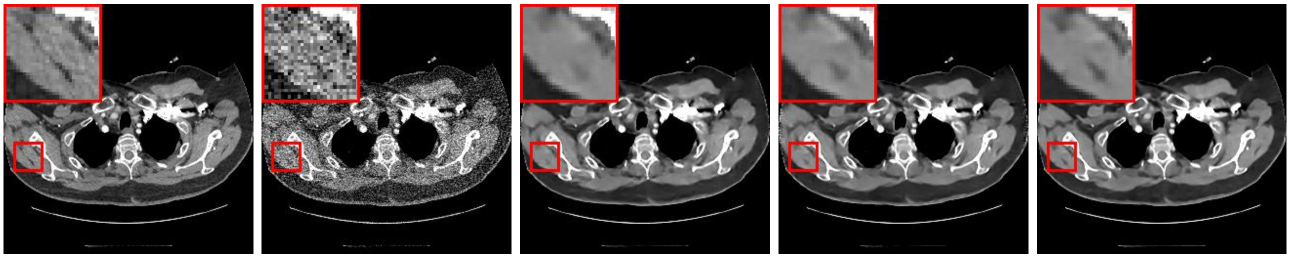
| | **Depth of conv. kernels** | | | | **Number of convolutions** | | |
|---|---|---|---|---|---|---|---|
| | **16** | **32** | **48** | **64** | **3** | **4** | **5** |
| PSNR (dB) | 44.79 | 45.30 | 46.10 | 46.21 | 45.21 | 46.10 | 46.03 |
| Number of parameters | 7,071 | 27,951 | 62,655 | 111,183 | 41,919 | 62,655 | 83,391 |
| Average testing time (s) | 0.153 | 0.189 | 0.239 | 0.278 | 0.201 | 0.239 | 0.282 |

### 4.2 Results on TCIA and NBIA test sets

We conducted extensive experiments on different algorithms. Besides the traditional FBP,[35] two state-of-the-art neural network based approaches got involved here for comparison, which are FBPConvNet[8] and RED-CNN.[20] The reconstruction quality was evaluated by the metrics Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). The experimental results of TCIA data on various dose levels are summarized in Table 2. The proposed LDA-LS returns the best results with the least number of parameters. Some representative testing images in NBIA dataset with dose level 2.5% are visualized in Figure 2 to give additional qualitative justification. It can be seen that LDA-LS preserves the detailed structures well and achieves the best quality.

Table 2. Quantitative results (Mean ± Standard Deviation) of the LDCT reconstructions of TCIA data of dose level 10%.

| **Dose Level** | $1.0 \times 10^5$ | | $5.0 \times 10^4$ | | $2.5 \times 10^4$ | | Parameters |
|---|---|---|---|---|---|---|---|
| | PSNR (dB) | SSIM | PSNR (dB) | SSIM | PSNR (dB) | SSIM | |
| FBP[35] | 38.56±0.71 | 0.9664±0.0050 | 35.82±0.77 | 0.9384±0.0094 | 32.99±0.80 | 0.8898±0.0163 | N/A |
| FBPConvNet[8] | 44.20±0.56 | 0.9941±0.0010 | 42.62±0.57 | 0.9918±0.0014 | 41.07±0.59 | 0.9886±0.0019 | $1.0 \cdot 10^7$ |
| RED-CNN[20] | 44.16±0.55 | 0.9939±0.0009 | 42.78±0.63 | 0.9920±0.0015 | 41.11±0.55 | 0.9887±0.0019 | $1.8 \cdot 10^6$ |
| **LDA-LS** | **47.36±0.66** | **0.9970±0.0006** | **44.65±0.63** | **0.9947±0.0009** | **43.37±0.65** | **0.9932±0.0011** | $\mathbf{6.2 \cdot 10^4}$ |



(a) Reference (b) FBP (32.02) (c) FBPConvNet(38.14 (d) RED-CNN (38.27) (e) LDA-LS (40.14)

Figure 2. Reconstructed CT images in NBIA data with display window [-160, 240] HU. PSNRs (dB) are in the parentheses.

## 5. CONCLUSION

This paper proposed the Learned Descent Algorithm with Line Search (LDA-LS) for LDCT reconstruction. By learning the nonsmooth nonconvex regularizer parameterized as the composition of $l_{2,1}$ norm and a neural network, LDA-LS can produce high-quality LDCT images with very low computational cost. The regularization learned by LDA-LS is more interpretable, and the convergence is guaranteed in sharp contrast to existing unrolling methods. Numerical results justify the promising efficiency and robustness of the proposed method.

## REFERENCES

[1] Hounsfield, G. N., "Computerized transverse axial scanning (tomography): Part 1. description of system," *The British Journal of Radiology* **46**(552), 1016–1022 (1973).

[2] Cormack, A. M., "Representation of a function by its line integrals, with some radiological applications. ii," *Journal of Applied Physics* **35**(10), 2908–2913 (1964).

[3] Jing Wang, Tianfang Li, Hongbing Lu, and Zhengrong Liang, "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography," *IEEE Transactions on Medical Imaging* **25**(10), 1272–1283 (2006).

[4] Tipnis, S. et al., "Iterative reconstruction in image space (iris) and lesion detection in abdominal ct," in [*Medical Imaging 2010: Physics of Medical Imaging*], **7622**, 76222K, International Society for Optics and Photonics (2010).

[5] Zheng, X., Ravishankar, S., Long, Y., and Fessler, J. A., "Pwls-ultra: An efficient clustering and learning-based approach for low-dose 3d ct image reconstruction," *IEEE transactions on medical imaging* **37**(6), 1498–1510 (2018).

[6] Zhang, Z., Liang, X., Dong, X., Xie, Y., and Cao, G., "A sparse-view ct reconstruction method based on combination of densenet and deconvolution," *IEEE transactions on medical imaging* **37**(6), 1407–1417 (2018).

[7] Hu, D. et al., "Hybrid-domain neural network processing for sparse-view ct reconstruction," *IEEE Transactions on Radiation and Plasma Medical Sciences* **5**(1), 88–98 (2021).

[8] Jin, K. H., McCann, M. T., Froustey, E., and Unser, M., "Deep convolutional neural network for inverse problems in imaging," *IEEE Transactions on Image Processing* **26**(9), 4509–4522 (2017).

[9] Xie, S. and Yang, T., "Artifact removal in sparse-angle ct based on feature fusion residual network," *IEEE Transactions on Radiation and Plasma Medical Sciences* **5**(2), 261–271 (2021).

[10] Lee, H., Lee, J., Kim, H., Cho, B., and Cho, S., "Deep-neural-network-based sinogram synthesis for sparse-view ct image reconstruction," *IEEE Transactions on Radiation and Plasma Medical Sciences* **3**(2), 109–119 (2018).

[11] Kang, E., Chang, W., Yoo, J., and Ye, J. C., "Deep convolutional framelet denosing for low-dose ct via wavelet residual network," *IEEE transactions on medical imaging* **37**(6), 1358–1369 (2018).

[12] Yang, Q. et al., "Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss," *IEEE transactions on medical imaging* **37**(6), 1348–1357 (2018).

[13] Chun, I. Y., Huang, Z., Lim, H., and Fessler, J., "Momentum-net: Fast and convergent iterative neural network for inverse problems," *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). pp. DOI: 10.1109/tpami.2020.3012955.

[14] Chun, I. Y., Zheng, X., Long, Y., and Fessler, J. A., "Bcd-net for low-dose ct reconstruction: Acceleration, convergence, and generalization," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 31–40, Springer (2019).

[15] Hammernik, K. et al., "Learning a variational network for reconstruction of accelerated mri data," *Magnetic resonance in medicine* **79**(6), 3055–3071 (2018).

[16] Sun, J., Li, H., Xu, Z., et al., "Deep admm-net for compressive sensing mri," in [*Advances in neural information processing systems*], 10–18 (2016).

[17] Zhang, J. and Ghanem, B., "Ista-net: Iterative shrinkage-thresholding algorithm inspired deep network for image compressive sensing," (2017).

[18] Chen, Y., Liu, H., Ye, X., and Zhang, Q., "Learnable descent algorithm for nonsmooth nonconvex image reconstruction," *arXiv preprint arXiv:2007.11245* (2020).

[19] Lee, H., Lee, J., and Cho, S., "View-interpolation of sparsely sampled sinogram using convolutional neural network," in [*Medical Imaging 2017: Image Processing*], **10133**, 1013328, International Society for Optics and Photonics (2017).

[20] Chen, H. et al., "Low-dose ct with a residual encoder-decoder convolutional neural network," *IEEE transactions on medical imaging* **36**(12), 2524–2535 (2017).

[21] Chen, H. et al., "Low-dose ct via convolutional neural network," *Biomedical optics express* **8**(2), 679–694 (2017).

[22] Kang, E., Min, J., and Ye, J. C., "A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction," *Medical physics* **44**(10), e360–e375 (2017).

[23] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*International Conference on Medical image computing and computer-assisted intervention*], 234–241, Springer (2015).

[24] Geyer, L. L. et al., "State of the art: iterative ct reconstruction techniques," *Radiology* **276**(2), 339–357 (2015).

[25] Chun, I. Y. and Fessler, J. A., "Convolutional dictionary learning: Acceleration and convergence," *IEEE Transactions on Image Processing* **27**(4), 1697–1712 (2017).

[26] Ye, S., Ravishankar, S., Long, Y., and Fessler, J. A., "Spultra: Low-dose ct image reconstruction with joint statistical and learned image models," *IEEE Transactions on Medical Imaging* **39**(3), 729–741 (2019).

[27] Ye, S., Long, Y., and Chun, I. Y., "Momentum-net for low-dose ct image reconstruction," *arXiv preprint arXiv:2002.12018* (2020).

[28] Chen, H. et al., "Learn: Learned experts' assessment-based reconstruction network for sparse-data ct," *IEEE transactions on medical imaging* **37**(6), 1333–1347 (2018).

[29] Adler, J. and Öktem, O., "Learned primal-dual reconstruction," *IEEE transactions on medical imaging* **37**(6), 1322–1332 (2018).

[30] Dumoulin, V. and Visin, F., "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285* (2016).

[31] Nesterov, Y., "Smooth minimization of non-smooth functions," *Mathematical programming* **103**(1), 127–152 (2005).

[32] Clarke, F. H., [*Optimization and nonsmooth analysis*], vol. 5, Siam (1990).

[33] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

[34] Glorot, X. and Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks," in [*In Proceedings of the International Conference on Artificial Intelligence and Statistics. Society for Artificial Intelligence and Statistics*], (2010).

[35] Kak, A. C., Slaney, M., and Wang, G., "Principles of computerized tomographic imaging," *Medical Physics* **29**(1), 107 (2002).

[36] Paszke, A. et al., "Pytorch: An imperative style, high-performance deep learning library," in [*Advances in Neural Information Processing Systems 32*], 8024–8035, Curran Associates, Inc. (2019).

[37] McCollough, C. et al., "Data from low dose ct image and projection data [data set]," *The Cancer Imaging Archive.* (2020).

[38] De Man, B. and Basu, S., "Distance-driven projection and backprojection," in [*2002 IEEE Nuclear Science Symposium Conference Record*], **3**, 1477–1480, IEEE (2002).

[39] De Man, B. and Basu, S., "Distance-driven projection and backprojection in three dimensions," *Physics in Medicine & Biology* **49**(11), 2463 (2004).

[40] li, T., Lu, H., and Liang, Z., "Penalized weighted least-squares approach to sinogram noise reduction and image reconstruction for low-dose x-ray computed tomography," *IEEE transactions on medical imaging* **25**, 1272–83 (11 2006).

[41] Niu, S. et al., "Sparse-view x-ray ct reconstruction via total generalized variation regularization," *Physics in Medicine & Biology* **59**(12), 2997 (2014).