The Practicality of Stochastic Optimization in Imaging Inverse Problems

The Practicality of Stochastic Optimization in Imaging Inverse Problems

Junqi Tang, *Member, IEEE*, Karen Egiazarian, *Fellow, IEEE*, Mohammad Golbabaee, *Member, IEEE* and Mike Davies, *Fellow, IEEE*

***Abstract*—In this work we investigate the practicality of stochastic gradient descent and its variants with variance-reduction techniques in imaging inverse problems. Such algorithms have been shown in the large-scale optimization and machine learning literature to have optimal complexity in theory, and to provide great improvement empirically over the deterministic gradient methods. However, in some tasks such as image deblurring, many of such methods fail to converge faster than the deterministic gradient methods, even in terms of epoch counts. We investigate this phenomenon and propose a theory-inspired mechanism for the practitioners to efficiently characterize whether it is beneficial for an inverse problem to be solved by stochastic optimization techniques or not. Using standard tools in numerical linear algebra, we derive conditions on the spectral structure of the inverse problem for being a suitable application of stochastic gradient methods. Particularly, if the Hessian matrix of an imaging inverse problem has a fast-decaying eigenspectrum, then our theory suggests that the stochastic gradient methods can be more advantageous than deterministic methods for solving such a problem. Our results also provide guidance on choosing appropriately the partition minibatch schemes, showing that a good minibatch scheme typically has relatively low correlation within each of the minibatches. Finally, we present numerical studies which validate our results.**

***Index Terms*—Imaging Inverse Problems, Stochastic Optimization, Large-scale Optimization.**

I. Introduction

STOCHASTIC gradient-based optimization algorithms have been ubiquitous in real-world applications which involve solving large-scale and high-dimensional optimization tasks, particularly in the field of machine learning [2]–[4], due to their scalability to the size of the optimization problems. In this work we study the practicality of stochastic gradient-based optimization algorithms in imaging inverse problems, which are also large-scale and high-dimensional by nature. The class of problems we consider, with typical examples including image deblurring, denoising, inpainting, superresolution, demosaicing, tomographic image reconstruction, etc, can be generally formulated as the following:

$$x^* \in \arg\min_{x\in\mathcal{X}} \left\{ F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x) \right\}, \qquad (1)$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex set and we denote by $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x) := \frac{1}{n}\sum_{i=1}^{n} \bar{f}(a_i, b_i, x)$ the data fidelity term, with a regularization term $g(x)$. We assume each $f_i(x) := \bar{f}(a_i, b_i, x)$ to be proper, convex and smooth. In the classical setting of supervised machine learning, the variable x contains the parameters of a classifier, while the vectors $\{a_1; a_2; ...; a_n\}$ represent the features of training data samples, and $\{b_1; b_2; ...; b_n\}$ denote the corresponding labels. In the imaging inverse problems we are interested in this work, they represent the vectorized image, the forward measurements and the observations, respectively.

To be more specific, we denote here a noisy linear measurement[1] model with a ground-truth vectorized image $x^\dagger$ which is to be estimated, an n by d matrix $A := [a_1; a_2; ...; a_n] \in \mathbb{R}^{n\times d}$ which denotes the measurement operator, additive noise denoted by vector $w \in \mathbb{R}^n$, and the noisy measurement data denoted by vector $b := [b_1; b_2; ...; b_n] \in \mathbb{R}^n$:

$$b = Ax^\dagger + w, \quad A \in \mathbb{R}^{n\times d}. \qquad (2)$$

A commonly-used data fidelity term in imaging inverse problems is the least-squares loss:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(a_i^T x - b_i)^2 = \frac{1}{2n}\|Ax - b\|_2^2, \qquad (3)$$

while we typically obtain a robust estimator of $x^\dagger$ via jointly minimizing the least-squares data-fidelity term with a structure-inducing regularization $g(\cdot)$ which encodes prior information we have regarding $x^\dagger$. Here we will assume $g(x)$ to be a proper convex and lower semi-continuous function and is possibly non-smooth. In imaging inverse problems, a commonly-used regularization is the sparsity-inducing norm penalty on either synthesis domain or analysis domain, with representative examples being the ℓ_1 regularization on wavelet coefficients, and the total-variation (TV) regularization [5].

Traditionally, imaging inverse problems are solved often by minimizing the regularized least-squares via deterministic first-order solvers, such as the proximal gradient descent [6], [7], its accelerated [8]–[11] and primal-dual variants [12], [13]. The

J.Tang and M.Davies are with the School of Engineering, University of Edinburgh; K.Egiazarian is with the Faculty of Information Technology and Communication Sciences, Tampere University; M.Golbabaee is with the Department of Computer Science, University of Bath. This work is supported by H2020-MSCA-ITN 642685 (MacSeNet) , ERC Advanced grant 694888, C-SENSE and a Royal Society Wolfson Research Merit Award. A preliminary version [1] of this work has been published as a conference paper in ICASSP 2019. Correspondence to J.Tang (Email: J.Tang@ed.ac.uk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes additional theoretical results. This material is 183KB in size.

[1]For the non-linear inverse problems, the measurement model is usually written as $b = A(x^\dagger) + w$ in the literature, with a non-linear mapping $A(\cdot) : \mathbb{R}^d \to \mathbb{R}^n$.

iterates of the proximal gradient descent with least-squares data-fidelity loss (3) can be written as:

$$\begin{aligned}&\textbf{Proximal gradient descent} - \text{Initialize } x^0 \in \mathcal{X}\\ &\text{For} \quad j = 0, 1, 2, ..., t\\ &\lfloor \; x^{j+1} = \text{prox}_{\eta g}[x^j - \eta \cdot \nabla f(x^j)]\end{aligned}$$

where we denote η as the step-size and for least-squares data-fidelity term $\nabla f(x^j) = \frac{1}{n} A^T(Ax^j - b)$. We denote the proximal operator as:

$$\text{prox}_{\eta g}(\cdot) = \arg\min_{x \in \mathcal{X}} \frac{1}{2}\|x - \cdot\|_2^2 + \eta g(x). \tag{4}$$

Unlike the deterministic gradient methods which need to compute a full gradient in each iteration, the stochastic gradient descent methods [2], [14] randomly select one or a few functions $f_i(x)$ in each iteration, compute an efficient unbiased estimate of the full gradient $\nabla f(x)$ and perform the descent step. When the composite optimization task (1) is large-scale and high-dimensional, stochastic gradient methods are able to achieve scalability and usually much more preferred than the deterministic gradient methods in machine learning applications.

In recent years, researchers have developed several advanced variants of stochastic gradient methods, namely, the variance-reduced stochastic gradient methods [15]–[18]. In each iteration of these stochastic algorithms, a more delicate stochastic gradient estimator is computed, which can reduce the variance of the stochastic gradient estimator progressively, with small computational overheads, and hence significantly improve the convergence rate of stochastic gradient methods. Most recently, by further combining the variance-reduction methods with the Nesterov's momentum acceleration technique which was originally designed to accelerate the deterministic gradient methods [19], researchers [20]–[22] have developed several accelerated stochastic gradient algorithms which can provably achieve the worse-case optimal convergence rate of gradient-based methods for (1).

While having been a proven success both in theory and in machine learning applications, there are few papers so far in the literature which demonstrate the performance benefit of the stochastic gradient methods in imaging applications. The most noticeable application of stochastic-gradient-type methods so far is the tomographic image reconstruction such as PET and CT [23]–[26]. The ordered-subsets algorithms [27]–[29], which are similar to SGD (the only difference is that they access the minibatches in a deterministic order which is prechosen judiciously tailored to the applications), are routinely used in clinical PET and CT systems for efficient reconstructions. The ordered-subsets methods have also been shown to have fast initial convergence rates in other applications such as multi-coil MRI reconstruction [30] and image restoration tasks [31]. It is worth noting that these ordered subsets/incremental gradient methods are not variance-reduced and require diminishing step-sizes to ensure convergence, which make them only to be able to converge fast initially at a low optimization accuracy regime, but significantly slowed down in later stage of optimization since the step-sizes shrink towards zero. Here we are focusing on the growing interest in importing stochastic gradient algorithms from machine learning [2]–[4]. Can stochastic gradient methods, especially the ones with variance-reduction techniques, significantly accelerate the solution of inverse problems as they did for machine learning? If not, why might stochastic optimization be inefficient for some inverse problems? How could we help practitioners to characterize whether a given inverse problem is suitable for stochastic gradient methods or not? This work is aimed at answering these questions in a systematic way.

A. Highlights of this work

We make the following contributions:

1) A metric for predicting stochastic acceleration: We start by a motivational analysis and propose to evaluate the limit of possible acceleration of a stochastic gradient method over its full gradient counterpart by measuring a metric which we call the *Stochastic Acceleration* (SA) factor, based on the ratio of the Lipschitz constants of the minibatch stochastic gradient and the full gradient. We also discover numerically that the SA factor is able to characterize the benefits of using randomized optimization techniques, and that not all imaging problems have a large SA factor.

2) Understanding the relationship between the structure of inverse problems and stochastic acceleration: We provide lower and upper bounds for the stochastic acceleration factors, for the linear imaging inverse problems where least-squares loss is applied as the data fidelity term. Our results suggest that:

If a linear inverse problem's Hessian matrix has an eigenspectrum which is fast-decaying, then it typically will have a large SA factor, which suggests that it can be characterized as a suitable application for stochastic gradient methods.

And vice-versa: if such an inverse problem's Hessian matrix has a slowly-decaying eigenspectrum, then it will have a small SA factor and can be deemed as unsuitable for stochastic gradient methods.

3) A measure of the comparative performance of different minibatch partitions: While the spectral properties of the forward operator fundamentally determine the suitability of stochastic gradient methods for an inverse problem, we show that in practice for some inverse problems, different choices of partition can lead to different convergence rates for stochastic gradient algorithms [24]. One of our lower bounds for SA factors suggests that:

If a partition scheme generates minibatches which have low local coherence structure, i.e. the measurements within minibatches have small correlation to each other, then it can be superior to other partition schemes which have high local coherence structure.

The SA factors and the lower bounds we propose are aimed to provide the practitioners with efficient ways to check whether they should use stochastic gradient techniques or classical deterministic gradient methods to solve a given

inverse problem, and also compare between different partition minibatch schemes and choose the best one among them in practice.

B. Related works

1) Proximal splitting schemes: The literature of proximal splitting algorithms in imaging is vast. In this work we mainly focus on the stochastic versus deterministic versions of the forward-backward splitting which is the most representative and widely-applied. Our analysis and numerical studies do not directly cover other splitting methods which are also popular in imaging inverse problems, such as the Douglas-Rachford splitting/ADMM [32], [33], block-coordinate descent methods [34], [35], and the variable-metric methods [36]–[38]. Nevertheless we believe that these classes of proximal splitting algorithms would also have similar stochastic acceleration limits.

2) Ordered-subsets methods in imaging: As mentioned before, in some medical imaging tasks such as CT and PET image reconstruction, a family of "SGD-like" minibatch gradient methods – namely the ordered-subset methods [27]–[29], are widely used in clinical practice due to its fast initial convergence compared to deterministic full gradient methods. The ordered-subsets methods take almost the same form of SGD with partition minibatches, except that they use a deterministic ordering of accessing the minibatches. Such an ordering is judiciously chosen for good empirical performance in tomography reconstruction [28], but the benefit of using deterministic ordering over random ordering has not yet been well-understood theoretically. Our current results do not cover such potential acceleration on certain inverse problems by designing deterministic minibatch orderings tailored to the specific tasks – this is an interesting future direction. On the other hand, current ordered-subset methods are not variance-reduced and hence can only provide fast initial convergence. It would be important to study whether these methods can be further improved by using variance-reduction techniques [16], [21], [39], to ensure fast global convergence.

3) Minibatch schemes in machine learning and imaging: In our work we focus on the partition minibatch schemes which best-suit the imaging practice [23]–[25], due to its implementation benefits. It is worth noting that, in stochastic optimization literature, this is not the only standard way of selecting minibatches [40], while different minibatch schemes suit different application scenarios. For instance, in machine learning we often do not use a fixed partition but reshuffle the data and generate new partitions in each epoch of SGD methods [41], [42], since the statistics of the datasets can be very different, and we generally do not know whether a partition is good or not. There is an important difference in the practical aspect of applying minibatch stochastic gradient methods in machine learning and imaging inverse problems. For most of the real-world imaging systems, for example the CT and PET, the measurements are fixed for the imaging devices, and we can pre-determine a good mini-batch partition beforehand [29], [43], and use the same partition each time we reconstruct an image. However for machine learning applications, if one wish to compute good minibatch partitions, it has to be done for different training datasets and different models [44], [45], which requires significant extra computations in practice.

Moreover, in some machine learning scenarios we may have imbalanced number of samples for different classes in training data, or there may exist a few out-of-distribution samples, that in such cases, a biased importance sampling would be beneficial for stochastic gradient methods since some data points are more important than others [46]–[48]. However, in imaging inverse problems, usually measurements are nearly equivalently important and the benefits of biasing the sampling distribution could be only incremental.

4) Plug-and-Play / Regularization-by-Denoising schemes: We also believe that our results can be extended for comparing the stochastic and deterministic versions of Plug-and-Play (PnP) [49]–[52] and Regularization-by-Denoising (RED) schemes [53]–[55] with advanced image priors based on applying image denoisers such as BM3D [56] and DnCNN [57] in a plug-in manner, since they have a similar algorithmic structure as the classical proximal splitting and gradient-based methods. We leave these as promising directions for future work.

C. Outline

Now we set out the rest of the paper. In section II we describe our notations and definitions which will be frequently used throughout the paper. We then present in section III a motivating example with a negative result of state-of-the-art stochastic gradient methods in a space-varying image deblurring task. Then in section IV, we provide a theoretical analysis regarding the limitation of stochastic optimization algorithms, and propose the theory-inspired SA factors. In section V, we focus on linear imaging inverse problems and present bounds for the SA factors with respect to the spectral properties of the forward operator, and hence derive a condition for an inverse problem to be a suitable application of stochastic gradient methods. Meanwhile, we also present results which link the SA factors with a local coherence property of partition minibatch schemes. In section VI, we present numerical experiments for the evaluation of our findings. Final remarks appear in section VII, while we include the proofs of our theoretical results in the appendix.

II. Notations and definitions

We now make clear some notations which will occur frequently throughout this paper. We denote an image $X \in \mathbb{R}^{d_1 \times d_2}$ in its vectorized (raster) form $x \in \mathbb{R}^d$ where $d = d_1 \times d_2$. Denote X's columns as $x_1, x_2, ..., x_{d_2} \in \mathbb{R}^{d_1}$, and $X = [x_1, x_2, ..., x_{d_2}]$, then $x = [x_1; x_2; ...; x_{d_2}]$. Without specification, the scalar n denotes the number of measurements, while d denotes the number of pixels, and m denotes the size of the minibatches, while K is the number of minibatches. For a positive integer q, the notation $[q]$ represents the collection of all positive integers up to q : $[1, ..., q]$. When we write $m = \frac{n}{K}$, we implicitly assume that $n \mod K = 0$ – this

is just for simplification of presentation, without the loss of generality.

For a given vector v and a scalar $p \geq 1$, we write its ℓ_p norm as $\|v\|_p$. We write the j-th row of A as a_j, and $A = [a_1; a_2; ...; a_n]$. We denote the transpose of A as A^T. We describe $\bar{\mathcal{I}} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$ as the partition of indices for a subsampling scheme, where $\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K = [n]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \forall i \neq j \in [K]$. Meanwhile, we use superscript indexing $S^1, S^2, ..., S^K$ to denote the corresponding row subsampling operators supported on the index set $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K$. For a given forward operator $A \in \mathbb{R}^{n\times d}$, we denote its spectral norm as $\|A\|$, and its Frobenius norm as $\|A\|_F$. We denote the k-th largest eigenvalue of a symmetric matrix $H \in \mathbb{R}^{d\times d}$ as $\sigma_k(H)$. We denote the $\ell_{1\to 2}$ inducing norm of A as:

$$\|A^T\|_{1\to 2}^2 := \max_{i\in[n]} \|a_i\|_2^2. \tag{5}$$

For a convex function $f(\cdot)$, we denote its convex conjugate as $f^*(\cdot)$. In this paper we consider the partition minibatch sampling which is the most widely-applied in practice:

Definition II.1 (Partition minibatch sampling): For a given minibatch index partition $\{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K\}$ where $\mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K = [n]$ and $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \forall i \neq j \in [n]$, the minibatches and the gradients are defined as the following:

$$\begin{aligned} f_{\mathcal{I}_k}(x) &= \frac{K}{n} \sum_{i\in\mathcal{I}_k} f_i(x), \\ \nabla f_{\mathcal{I}_k}(x) &:= \frac{K}{n} \sum_{i\in\mathcal{I}_k} \nabla f_i(x), \quad k \in [K]. \end{aligned} \tag{6}$$

We will then consider problems with the following standard smoothness (gradient-Lipschitz) conditions on the full batch $f(x)$ and minibatches $f_{\mathcal{I}_k}(x)$:

Definition II.2: (Smoothness of the Full-Batch and the Mini-Batches.) $f(\cdot)$ is $L_{\nabla f}$-smooth and each $f_{\mathcal{I}_k}(\cdot)$ is L_b-smooth, that is:

$$f(x) - f(y) - \nabla f(y)^T(x-y) \leq \frac{L_{\nabla f}}{2}\|x-y\|_2^2, \quad \forall x, y \in \mathcal{X}, \tag{7}$$

and L_b being the smallest constant such that for all $\mathcal{I}_k \in \bar{\mathcal{I}}$,

$$f_{\mathcal{I}_k}(x) - f_{\mathcal{I}_k}(y) - \nabla f_{\mathcal{I}_k}(y)^T(x-y) \leq \frac{L_b}{2}\|x-y\|_2^2, \forall x, y \in \mathcal{X}. \tag{8}$$

The smoothness of a function $f(\cdot)$ essentially means that any function value $f(x)$ is upper-bounded by a quadratic approximation $f(y)+\nabla f(y)^T(x-y)+\frac{L_{\nabla f}}{2}\|x-y\|_2^2$. For least-square loss $f(x) = \frac{1}{2n}\|Ax - y\|$, we have $L_{\nabla f} = \frac{\|A\|_2^2}{n}$. The smoothness of $f(\cdot)$ and $f_{\mathcal{I}_k}(\cdot)$ also means that their gradients are Lipschitz-continuous [58], that is:

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L_{\nabla f}\|x-y\|_2 \quad \forall x, y \in \mathcal{X}, \tag{9}$$

and also:

$$\|\nabla f_{\mathcal{I}_k}(x) - \nabla f_{\mathcal{I}_k}(y)\|_2 \leq L_b\|x-y\|_2 \quad \forall x, y \in \mathcal{X}. \tag{10}$$

III. A MOTIVATING EXAMPLE

Image deblurring is an important type of imaging inverse problems and has been studied intensely during the recent decades. For uniform deblurring, due to the cyclic structure of the deconvolution, FFT-based ADMM[2] variants have been shown to be remarkably efficient [59] when compared to classic gradient-based solvers such as FISTA [8]. Such techniques, although being computationally efficient, are specifically tailored to a restricted range of problems where the observation models are diagonalizable by a DFT. For image deblurring, it is often not realistic to assume that imaging devices induce a uniform blur [60]. If the blurring is different across the image, then the efficient implementation of ADMM is not effective in general. In such cases, the standard ADMM and deterministic gradient methods such as FISTA can be computationally expensive since the deblurring problems we have in practice are usually large-scale and high-dimensional. It is therefore natural to ask: can stochastic gradient methods offer us a more efficient solution?

We start by a simple space-varying deblurring [60] example where a part (sized 256 by 256) of the "Kodim04" image from *Kodak Lossless True Color Image Suite* [61] is blurred with a space-varying blur kernel which imposes less blurring at the center but increasingly severe blurring towards the edge. For the shape of the blur kernel, we choose the out-of-focus kernel provided in [62]. We also add a small amount of noise to the blurred image.

We test the effectiveness of several algorithms by solving the same TV-regularized least-squares problem, with the regularizer $g(x) = \lambda\|Dx\|_1$ (where D is the 2D differential operator and λ is the regularization parameter), to get an estimation of the ground truth image. The proximal operator of TV semi-norm is defined as:

$$\text{prox}_{TV}(z) := \arg\min_x \frac{1}{2}\|x-z\|_2^2 + \lambda\|Dx\|_1, \tag{11}$$

which does not have a closed-form solution. The most common approach to approximate this is to run accelerated gradient descent (AGD) on the dual of (11), as proposed in [9]. Here we run 10-iterations of AGD which initialize with z for warm-start, following the standard implementation of UnLocbox package [63].

The algorithms we test in the experiments include the accelerated full gradient method FISTA [9], proximal SGD [64], the proximal SVRG [17], [65] and its accelerated variant, Katyusha algorithm [21] which has achieved optimal convergence rate in theory for (1). These methods we choose for comparison here are representative examples for different classes of algorithms. FISTA is one of the most widely-applied deterministic proximal gradient methods for imaging inverse problems. The proximal SGD is the basic form of stochastic gradient method for (1), while the proximal SVRG equip the vanilla SGD with variance-reduction techniques and achieve

[2]The computationally demanding sub-problems of *alternating direction method of multipliers* (ADMM) in this case can be solved with an efficient matrix inversion by FFT due to the cyclic structure of the uniform deconvolution.

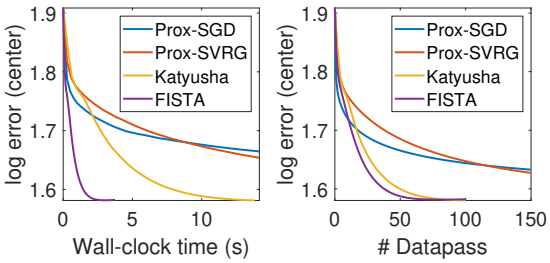


Fig. 1: *The estimation error plot for the deblurring experiment. The plots correspond to the estimation error of the central part (226 by 226) of the image.*

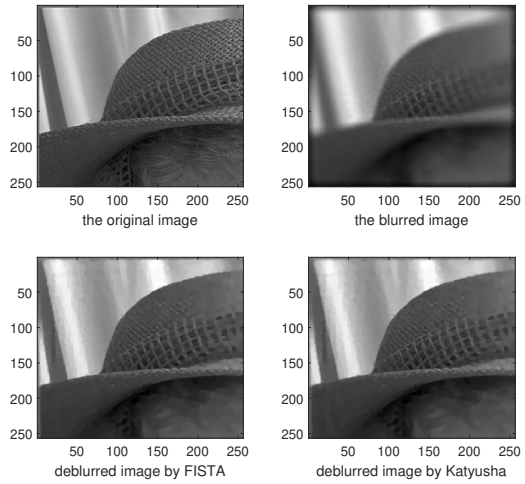


Fig. 2: *Up-left: the orignal image; up-right: the blurred image which is also corrupted with Gaussian noise; Down-left: deblurred image by FISTA; Down-right: deblurred image by Katyusha algorithm.*

algorithmic improvements both in theory and in machine learning applications. The Katyusha algorithm is an accelerated proximal SVRG method using the momentum technique and represents the state-of-the-art stochastic gradient methods for solving (1). We use the checkerboard partitioning [31] to generate 16 minibatches (which is a 4×4 2D-interleaving) for the stochastic gradient methods hence the minibatch size is 4096. On this experiment we report a negative result in Fig.1 for all these randomized algorithms. The most efficient solver in this task is the full gradient method FISTA in terms of wall-clock time and number of epochs (datapasses). Note that the epoch count provides a measure of the computation of the gradients *without* the cost of the proximal operators. We will subsequently use this when we wish to focus solely on the gradient costs in different algorithms. The minibatch size was chosen to maximize the performance of the stochastic methods.

Note that while the stochastic gradient methods do not offer in this example any benefit in terms of epoch counts as suggested by their theories, the actual wall-clock time performance-gap between them and FISTA is even larger. This is because of the computational overhead of proximal operators – the stochastic proximal gradient methods need to compute multiple times the total-variation proximal operators within each epoch. Hence when evaluating the benefits of stochastic gradient methods in imaging problems we should note that any benefits observed in terms of epoch count may be over optimistic given the non-negligible computational overhead of the proximal operators.

IV. Limitations of stochastic optimization

The previous deblurring example appears to be contrary to the popular belief among the stochastic optimization community and the experience of machine learning practitioners, that stochastic gradient methods are much faster in terms of gradient complexity measured by the epoch counts than deterministic gradient methods in solving large scale problems. To be specific – to achieve an objective gap suboptimality of $F(x) - F(x^\star) \leq \varepsilon$, optimal stochastic gradient methods needs only $\Theta\left(n + \sqrt{nL/\varepsilon}\right)$ evaluations of ∇f_i where $L = \max_i L_i$ and L_i denotes the smallest Lipschitz constant of ∇f_i, see e.g. [20], [21], while $\Theta\left(n\sqrt{L/\varepsilon}\right)$ are needed for optimal full gradient methods [66]. Where is the loophole?

It is often easily ignored that the complexity results above are derived under different smoothness assumptions. For convergence bounds of the full gradient, the full smooth part of the cost function's gradient $\nabla f(.)$ is assumed to be L-Lipschitz continuous, while for the case of stochastic gradient, every individual function's gradient $\nabla f_i(.)$ is assumed to have a Lipschitz constant L. Now we can clearly see the subtlety: to compare these complexity results and make meaningful conclusions, one has to assume that these two Lipschitz constants are roughly the same. While this can be true, and is true for many problems, there are exceptions – image deblurring is one of them.

We illustrate here some extreme examples for the two smoothness constants to demonstrate this possible dramatic difference:

Let $f(x) = \frac{1}{2n}\|Ax - b\|_2^2 = \frac{1}{n}\sum_{i=1}^n \frac{1}{2}(a_i^T x - b_i)^2 := \frac{1}{n}\sum_{i=1}^n f_i(x)$.

(1) If $a_1 = a_2 = a_3 ..., = a_{n-1} = a_n$, then $L_{\nabla f} = L_b$.

(2) If $A = I$, then $L_{\nabla f} = \frac{1}{K} L_b$.

A. A motivational analysis

In order to identify the potential of a certain imaging inverse problem to be more efficiently solved using stochastic gradient methods, we start by deriving a motivating theorem comparing the convergence rates in terms of epoch counts for the optimal full gradient methods as well as the optimal stochastic gradient methods. Recall that in our motivational example, the performance gap in terms of actual running time between deterministic and stochastic gradient methods is larger than their performance gap measured purely by epoch counts due to the computational overhead of proximal operators. In our motivational analysis, we consider the unregularized case where $g(\cdot) = 0$ to focus on the fundamental comparison of the iteration complexity of these algorithms measured by the epoch counts. We also have additional results for the case where convex constraints (such as the total-variation constraint

[67]) are used as regularization, and we include them in the supplemental material.

We compare two classes of algorithms: the optimal deterministic gradient methods which meet the deterministic gradient-complexity lower bound [68, Theorem 3] and the optimal stochastic gradient methods which are able to match the stochastic gradient-complexity lower bound [68, Theorem 7]. The FISTA algorithm and the Katyusha algorithm are typical instances from these two classes of algorithms.

Definition IV.1: (The class of optimal deterministic gradient algorithms.) A deterministic gradient method $\mathcal{A}_{\text{full}}$ is called optimal if for any $t \geq 1$, the update of the t-th iteration $x^t_{\mathcal{A}_{\text{full}}}$ satisfies:

$$F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star) \leq \frac{C_1 L_{\triangledown f}\|x^0 - x^\star\|_2^2}{t^2}, \tag{12}$$

for some positive constant C_1.
It is known that the original FISTA algorithm satisfies this definition with $C_1 = 2$ [9], while recent works [69], [70] show that the constant can be further improved to $C_1 = 1$ via optimizing the momentum parameter. We also define the class for optimal stochastic gradient methods:

Definition IV.2: (The class of optimal stochastic gradient algorithms.) A stochastic gradient method $\mathcal{A}_{\text{stoc}}$ is called optimal if for any $t \geq 1$ and $K \geq 1$, after a number of $t{\cdot}K$ stochastic gradient evaluations, the output of the algorithm $x^t_{\mathcal{A}_{\text{stoc}}}$ satisfies:

$$\begin{aligned}&\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)\\ &\leq \frac{C_2[F(x^0) - F(x^\star)]}{t^2} + \frac{C_3 L_b\|x^0 - x^\star\|_2^2}{Kt^2},\end{aligned} \tag{13}$$

for some positive constants C_2 and C_3.
Note that the accelerated stochastic variance-reduced gradient methods [20]–[22] satisfy this definition with different constants of C_2 and C_3.

Now we are ready to present the motivational theorem, which follows from combining the existing convergence results of the lower bounds for the stochastic and deterministic first-order optimization [58], [68].

Theorem IV.3: Let $g(\cdot) = 0$ and $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2^2 \leq 1\right\}$. Denote an optimal deterministic algorithm $\mathcal{A}_{\text{full}}$ which satisfies Def. IV.1, and an optimal stochastic gradient algorithm $\mathcal{A}_{\text{stoc}}$ which satisifes Def. IV.2. For a sufficiently large dimension d, there exist two different sets of convex and L_b-smooth functions f_i, such that $F = \frac{1}{K}\sum_{i=1}^K f_i$ satisfies the following bounds respectively:

$$\frac{\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)}{F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star)} \geq \frac{c_0 L_b}{K L_{\triangledown f}}, \tag{14}$$

and,

$$\frac{\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)}{F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star)} \leq \frac{c_1 L_b}{K L_{\triangledown f}} + c_2, \tag{15}$$

with universal positive constants c_0, c_1 and c_2 which do not depend on L_b, $L_{\triangledown f}$, t and K.

We provide the proof in Appendix A. From this theorem we can see that with the same epoch count, the ratio of the objective-gap sub-optimality achieved by $\mathcal{A}_{\text{full}}$ and $\mathcal{A}_{\text{stoc}}$ can be upper and lower bounded by $\Theta(\frac{L_b}{KL_{\triangledown f}})$ in the worst case. In other words, for a fixed number of epochs, there exists a smooth finite-sum objective function, such that no optimal stochastic gradient method can achieve a speed-up more than $c_0 \cdot \frac{L_b}{KL_{\triangledown f}}$ times over any optimal deterministic gradient algorithm on minimizing this objective. Meanwhile, there also exist a smooth finite-sum objective function, such that no optimal deterministic gradient method can achieve a speed-up more than $c_1 \cdot \frac{L_b}{KL_{\triangledown f}} + c_2$ times over any optimal stochastic gradient methods. Motivated by the theory, we now further investigate and propose to evaluate the potential of stochastic acceleration simply by the ratio $\frac{KL_{\triangledown f}}{L_b}$ which dominates our upper and lower bound in Theorem IV.3.

B. Evaluating the limitation of SGD-type algorithms

We propose a metric called the *Stochastic Acceleration* (SA) factor based on our theoretical analysis in the previous section, in order to provide a way of characterizing whether for a given inverse problem and a certain partition minibatch sampling scheme, stochastic gradient methods should be preferred over the deterministic full gradient methods or not.

Definition IV.4: For a given disjoint partition minibatch index $[n] = \mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K := \bar{\mathcal{I}}$, where $\mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \forall i \neq j \in [K]$, with corresponding subsampling operators $[S^1, ...S^K]$, the Stochastic Acceleration (SA) factor is defined as:

$$\Upsilon(A, \bar{\mathcal{I}}, K) = \frac{KL_{\triangledown f}}{L_b}. \tag{16}$$

Note that while the SA factor is always greater than or equal to 1, we remind the reader that this is only a qualitative measure as we have discounted the constants in Theorem IV.3 and the computational cost of the proximal operator.

We next evaluate the SA factors for the least squares loss function $f(x) = \frac{1}{2n}\|Ax - b\|_2^2$ with different types of forward operators. We partition the data into minibatches and have:

$$f(x) = \frac{1}{2n}\|Ax - b\|_2^2 = \frac{1}{K}\sum_{k=1}^K f_{\mathcal{I}_k}(x), \tag{17}$$

where,

$$f_{\mathcal{I}_k}(x) := \frac{K}{2n}\|S^k Ax - S^k b\|_2^2, \tag{18}$$

The examples of forward operator A we consider here include the space-varying deblurring ($A_{\text{blur}} \in \mathbb{R}^{262144\times 262144}$), a random compressed sensing matrix with i.i.d Guassian random entries (with a size $A_{\text{rand}} \in \mathbb{R}^{500\times 2000}$), and a fan beam X-ray CT operator ($A_{\text{CT}} \in \mathbb{R}^{91200\times 65536}$). Meanwhile, in order to contrast with the application of stochastic gradient algorithms in machine learning, we also consider linear regression problems on two machine learning datasets: RCV1 dataset ($A_{\text{rcv1}} \in \mathbb{R}^{20242\times 47236}$), and Magic04 ($A_{\text{magic04}} \in \mathbb{R}^{19000\times 50}$).

For the X-ray CT image reconstruction example and deblurring example we use TV regularization for $g(x)$ in (1), while for the rest of the examples we use ℓ_1 regularization. We vectorize the image precisely as described in section II. The data-partition we select for the deblurring example is the

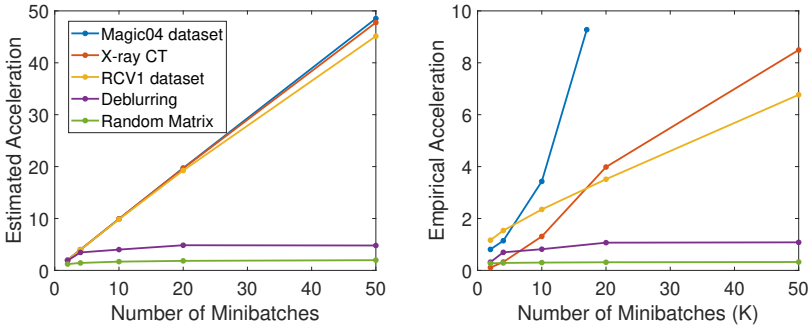


Fig. 3: *Left: Stochastic Acceleration (SA) function of inverse problems with different forward operators. Right: Empirical observation comparing the convergence of Katyusha and FISTA algorithm in 15 epochs.*

checkerboard partition proposed by [31], which was observed to be the most appropriate for image-restoration tasks:

$$f_{\mathcal{I}_k}(x) = \frac{K}{n} \sum_{i=1}^{\lfloor d_1/k_1 \rfloor} \sum_{j=1}^{\lfloor d_2/k_2 \rfloor} f_{(i-1)k_1 d_2 + (j-1)k_2 + v(k)}(x), \quad (19)$$

where $K = k_1 k_2$ and $v(k) = \mod(k, k_2) + d_2 \lfloor k/k_2 \rfloor$. For CT we use the standard view-based subsets where we interleave the projections from equally-spaced angles:

$$f_{\mathcal{I}_k}(x) := \frac{K}{n'} \sum_{i=1}^{\lfloor n'/K \rfloor} f_{k+iK}(x) \quad (20)$$

$$= \frac{K}{2n'} \sum_{i=1}^{\lfloor n'/K \rfloor} \sum_{j=m(k+iK)+1}^{m(k+iK+1)} (a_j^T x - b_j) \quad (21)$$

where we denote n' as number of views and m as the number of X-ray sensors in the CT system. The data-partition we choose for the rest of examples is the interleaved sampling, where the k-th minibatch is formed as the following:

$$f_{\mathcal{I}_k}(x) := \frac{K}{n} \sum_{i=1}^{\lfloor n/K \rfloor} f_{k+iK}(x) \quad (22)$$

$$= \frac{K}{2n} \sum_{i=1}^{\lfloor n/K \rfloor} (a_{k+iK}^T x - b_{k+iK}) \quad (23)$$

In Figure 3(a), we plot the SA factors for these 5 problem instances as a function of the number of minibatches along with the empirical acceleration observed when solving these problems. From the result demonstrated in the Figure 3 we find that indeed the stochastic methods have a limitation on some problems like deblurring and compressed sensing inverse problems with Gaussian random design matrices, where we see that the curve for the SA factor of such problems stays low and flat even when we increase the number of minibatches. For the machine learning datasets and X-ray CT imaging, the SA factor increases rapidly and almost linearly as we increase the number of minibatches, which is in line with observations in machine learning on the superiority of SGD and also the observation in CT image reconstruction of the benefits of using the ordered-subset methods [27], [29] which are similar to stochastic gradient methods.

The curves for the SA factor in Figure 3(a) qualitatively predict the empirical comparison result[3] of the Katyusha and FISTA algorithms shown on the Figure 3(b), where we observe that Katyusha offers no acceleration over the FISTA on either the deblurring or the compressed sensing inverse problem we have considered, but significantly outperforms FISTA on the other cases. Indeed, positive results for applying SGD-type algorithms on these problems are well-known already [2], [17], [27]. These results suggest that the SA factor we propose could be useful in characterizing whether an inverse problem is inherently a suitable candidate for stochastic gradient methods.

V. Local coherence structure, eigenspectrum, and stochastic acceleration

We now go deeper to investigate the relationship of the SA factor and the structure of the forward operator of the inverse problem. We restrict our analysis in this section to the cases where least-squares loss is used as the data-fidelity term.

Subsequently we will assume that each partition has an equal size m for the simplicity of presentation. We will find the following definition of the *local-accumulated-coherence* to be useful.

Definition V.1 (Local-Accumulated-Coherence): Give a partition $[n] = \mathcal{I}_1 \cup \mathcal{I}_2 \cup ... \cup \mathcal{I}_K := \bar{\mathcal{I}}$ for $A = [a_1; a_2; ...; a_n]$, the local-accumulated-coherence is defined as:

$$\mu_\ell(A, \bar{\mathcal{I}}, K) = \max_{q \in [K]} \max_{j \in \mathcal{I}_q} \sum_{k \in \mathcal{I}_q} |\langle a_j, a_k \rangle|. \quad (24)$$

The local-accumulated-coherence captures the correlation characteristic between the linear measurements within each partitioned minibatches[4]. As we will see, if a partition has a smaller local accumulated coherence than another partition, then it typically can have a better SA factor. This suggests that a good partition scheme should ensure that within each of the minibatches, the measurements are spread across the image space as uniformly as possible. Our theoretical result confirms the numerical observations in empirical works regarding the

[3]We compare the objective-gap convergence $(F(x^t) - F(x^\star))$ of FISTA and Katyusha for a fixed number of datapasses (epochs).

[4]Note that our definition of local accumulated coherence should not be confused with the definition of cumulative coherence in [71] which is regarding the column elements of a dictionary.

ordered-subsets methods [27], [31], [72], that it is often more desirable for a minibatch scheme to generate "balanced" minibatches which have nearly equivalent pixel activity [72].

The Gersgorin disk theorem [73] will be useful in our analysis, which relates a square symmetric matrix's eigenvalues with its entries, and links to the gradient-Lipschitz constant L_b – which in the least-squares context can be written as:

$$L_b = \frac{K}{n} \max_{k\in[K]} \|S^k A (S^k A)^T\| \tag{25}$$

in linear inverse problems. With this relationship we can lower-bound the SA factors with the ratio of $\frac{\|A\|^2}{\mu_\ell}$.

Theorem V.2 (Lower bounds for $\Upsilon(A,\bar{\mathcal{I}},K)$): The SA factor for any linear inverse problem with $f(x) = \frac{1}{2n}\|Ax-b\|_2^2$ is lower bounded as:

$$\Upsilon(A,\bar{\mathcal{I}},K) \geq \alpha_\ell(A,\bar{\mathcal{I}},K) := \frac{\|A\|^2}{\mu_\ell(A,\bar{\mathcal{I}},K)} \tag{26}$$

$$\geq \alpha_u(A,K) := \frac{K\|A\|^2}{n\|A^T\|_{1\to 2}^2} \tag{27}$$

$$\geq \alpha_s(A,K) := \frac{K\cdot\sigma_1(A^TA)}{\rho\cdot\sum_{i=1}^d \sigma_i(A^TA)}, \tag{28}$$

where:

$$\rho := \frac{\max_{i\in[n]}\|a_i\|_2^2}{\frac{1}{n}\sum_{j=1}^n \|a_j\|_2^2}. \tag{29}$$

We provide the proof in Appendix B. Note that most inverse problems we encounter usually satisfy (29) with $\rho = O(1)$. This is due to the fact that unlike machine learning applications where we may often have outliers in datasets, most imaging systems are well-designed with measurements having similar ℓ_2 norms. The second and the third inequalities are partition-independent and reveal a strong relationship between the SA factor and the spectral properties of the forward operator. The third inequality in Theorem V.2 suggests that, *if a linear inverse problem which satisfies (29) with $\rho = O(1)$ has a Hessian with a fast-decaying eigenspectrum, it typically will have a good SA factor*. They are also tight bounds if we do not impose additional structural assumptions on the forward operator – if A has identical rows we have $\Upsilon(A,\bar{\mathcal{I}},K) = \frac{K\|A\|^2}{n\|A^T\|_{1\to 2}^2} = K$, noting that $\|A^T\|_{1\to 2}^2 = \max_{i\in[n]}\|a_i\|_2^2$. If $\rho = 1$ which means that rows of A have the same ℓ_2 norm, we also have $\alpha_u(A,K) = \alpha_s(A,K)$.

We have derived partition independent lower bounds $\alpha_u(A,K)$ and $\alpha_s(A,K)$ which link the SA factor $\Upsilon(A,\bar{\mathcal{I}},K)$ with spectral properties of the forward operator. For some inverse problems which admit inferior partitions, these may be crude lower bounds since they should cover the worse case of partition choice. It is therefore insightful to derive a lower bound for the case where we randomly partition the data, which would enable us to have potentially better lower bound estimate for the SA factors. We provide the following lower bound using the Matrix Chernoff inequality and the union bound, following a similar argument by [74, Proposition 3.3]. We present the proof in Appendix D.

Theorem V.3 (Lower bounds for $\Upsilon(A,\bar{\mathcal{I}},K)$ for a random partition): If $\bar{\mathcal{I}}$ is a uniform random partition, then for $K \in \left[\frac{\|A\|^2}{\|A^T\|_{1\to 2}^2}, \min(n,d)\right]$, the following lower bounds hold with probability at least: $1 - d^2\left(\frac{e}{\delta}\right)^\delta$:

$$\Upsilon(A,\bar{\mathcal{I}},K) \geq \alpha_r(A,K,\delta) \geq \alpha_\sigma(A,K,\delta), \tag{30}$$

where,

$$\begin{aligned}\alpha_r(A,K,\delta) &:= \frac{1}{\frac{1}{K} + \delta\cdot\frac{\|A^T\|_{1\to 2}^2}{\|A\|^2}},\\ \alpha_\sigma(A,K,\delta) &:= \frac{1}{\frac{1}{K} + \delta\cdot\frac{\rho}{n}\cdot\frac{\sum_{i=1}^d\sigma_i(A^TA)}{\sigma_1(A^TA)}}.\end{aligned} \tag{31}$$

These lower bounds for a random partition scheme again demonstrates the strong relationship between the SA factor and the ratio $\frac{\|A^T\|_{1\to 2}^2}{\|A\|^2}$ which is controlled by the eigenspectrum $\frac{\sum_{i=1}^d\sigma_i(A^TA)}{\sigma_1(A^TA)}$. Note that due to the Matrix Chernoff inequality [75], this theorem holds with a probability $1-d^2\left(\frac{e}{\delta}\right)^\delta$, which is dimension-dependent, hence in theory the parameter δ needs to be sufficiently large for this bound to hold with high probability.

Meanwhile, we can also have an upper bound for the SA factor, independent of the partition $\bar{\mathcal{I}}$, in terms of the eigenspectrum of the Hessian matrix A^TA. This upper bound can be derived from a standard result [73, Theorem 4.3.15] using the fact that the matrix $(S^kA)^TS^kA$ and $S^kA(S^kA)^T$ share the same non-zero eigenvalues.

Theorem V.4 (Upper bound for $\Upsilon(A,\bar{\mathcal{I}},K)$): The SA factor for any linear inverse problem with $f(x) = \frac{1}{2n}\|Ax-b\|_2^2$ is upper bounded as:

$$\Upsilon(A,\bar{\mathcal{I}},K) \leq \beta(A,K) := \frac{\sigma_1\left(A^TA\right)}{\sigma_{\lfloor n-\frac{n}{K}+1\rfloor}\left(A^TA\right)}, \tag{32}$$

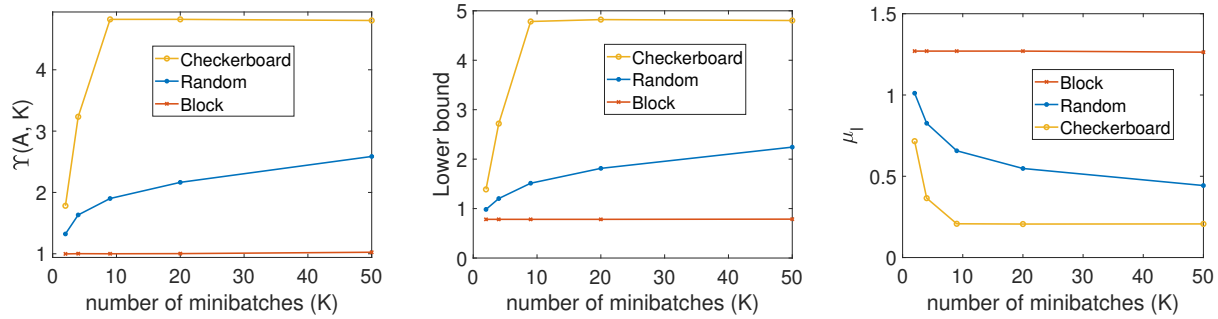
for any possible partition $\bar{\mathcal{I}}$.

We include the proof in Appendix C for completeness. The upper bound (32) suggests that, if the Hessian matrix A^TA has slowly-decaying eigenvalues at the tail, it indeed typically cannot have a large SA factor, no-matter how delicately we partition the forward operator A. The upper bound and the lower-bounds jointly suggest that, having a fast-decaying eigenspectrum of the Hessian is crucial for an inverse problem to have good SA factors.

VI. Numerical Evaluation

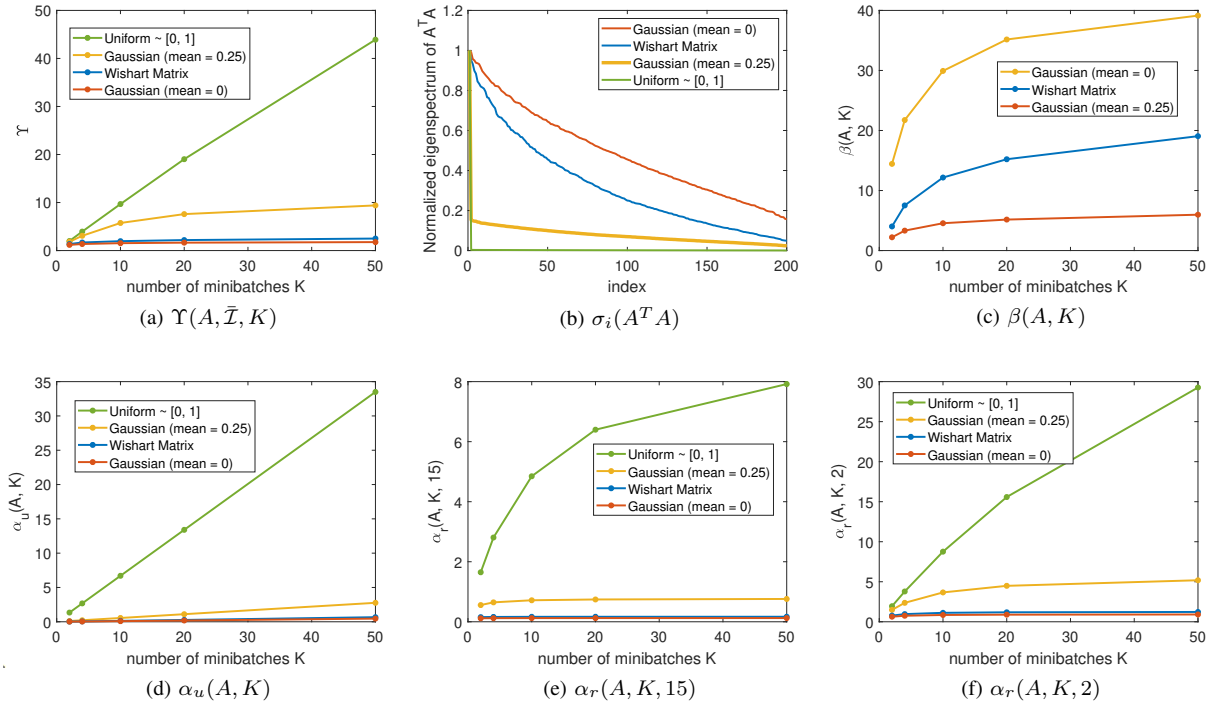
In this section, we design numerical experiments to validate our theoretical findings in the previous section, regarding how the SA factor depends on the inherent spectral structure of the forward operator, as well as the minibatch partition schemes we choose. We validate these theoretical results on various examples of linear inverse problems with different properties. We use a machine with 1.6 GB RAM, 1.80 GHz Intel Core i7-8550 CPU and MATLAB R2020a.

We start by the lower bound $\alpha_\ell(A,\bar{\mathcal{I}},K)$ we have presented in Theorem V.2 which suggests that, for some inverse problems, judiciously choosing the partition for minibatches is important – good choices of partitioning can have small local coherence and hence lead to larger SA factors in practice. In



(a) $\Upsilon(A, \bar{\mathcal{I}}, K)$ (b) $\alpha_\ell(A, \bar{\mathcal{I}}, K)$ (c) $\mu_\ell(A, \bar{\mathcal{I}}, K)$

Fig. 4: *(a) Stochastic Acceleration (SA) function under different partition choices for space-varying deblurring task. (b) lower bound estimate of SA factors via local accumulated coherence. (c) local accumulated coherence for each partition scheme.*



(a) $\Upsilon(A, \bar{\mathcal{I}}, K)$ (b) $\sigma_i(A^T A)$ (c) $\beta(A, K)$

(d) $\alpha_u(A, K)$ (e) $\alpha_r(A, K, 15)$ (f) $\alpha_r(A, K, 2)$

Fig. 5: *Stochastic Acceleration (SA) factors, eigenspectrum, and lower/upper bound estimates for forward operators (random matrices sized 200 by 1000) with different distributions, using random partitioning.*

Figure 4, we test three different partition schemes for our running example of space-varying deblurring: the checkerboard 2D-interleaving partitioning which we have described before, the random partitioning where we generate the partition index randomly without replacement, and the consecutive block partitioning[5], where we directly partition the forward operator A into K consecutive blocks and form the minibatches. The interleaving partitioning in this case provides the smallest local coherence, and hence its SA factors are the largest. While consecutive block partitioning leads to the largest local coherence, it offers no stochastic acceleration at all. The lower bound estimate $\alpha_\ell(A, \bar{\mathcal{I}}, K)$ of SA factors are actually very accurate in this case, as shown in the Figure 4(b).

In Figure 5 we present a simulation result where we generate 4 compressed sensing random design matrices of the same size $n = 200, d = 1000$ with different distributions and check the relationship of their SA factors and the eigenspectrum of their Hessian matrices. The forward operators we generate are:

(1) random Gaussian matrix with each entry drawn from a Gaussian distribution with zero-mean and unit-variance;

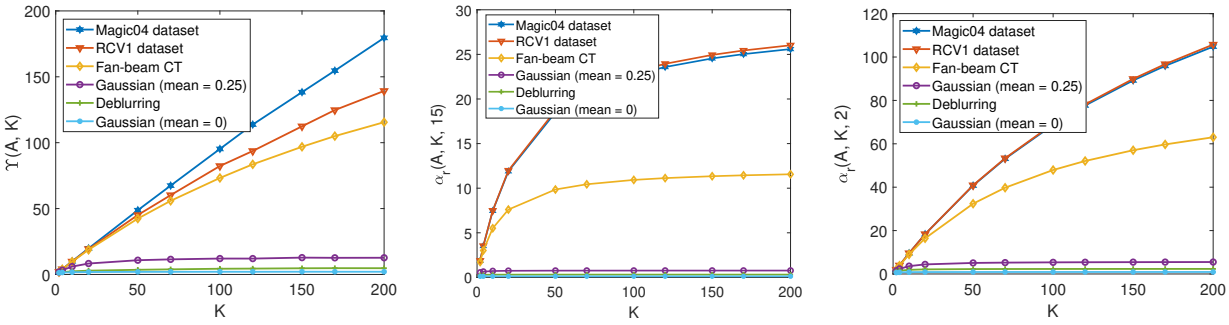
(2) subsampled Wishart matrix;

(3) random Gaussian matrix with each entry drawn from a Gaussian distribution with 0.25-mean and unit-variance;

(4) random matrix with each entry drawn from a uniform distribution supported on the interval $[0, 1]$.

From the experimental result we can observe that, these four forward operators have very different decay-rates on their

[5]which is basically $\mathcal{I}_k := [m(k-1)+1, m(k-1)+2, m(k-1)+3, ..., mk-1, mk], \forall k \in [K]$.



(a) $\Upsilon(A, \bar{\mathcal{I}}, K)$ (b) $\alpha_r(A, K, 15)$ (c) $\alpha_r(A, K, 2)$

Fig. 6: *SA factors, and lower bound estimates of inverse problems with different forward operators for random partition minibatches.*

Hessians' eigenspectrum, and correspondingly, very different SA factors. The case (4) has the fastest decay-rate, and has the largest SA factors and it grows almost linearly as the number of minibatches increases. The case (1) has the slowest decay-rate on the eigenspectrum, and correspondingly, it has the worst SA factors among the 4 cases. This numerical result is in broad agreement with our analysis.

Next we also test our lower bound $\alpha_r(A, K, \delta)$ for all the examples we have considered. We first compute the lower bound estimate $\alpha_r(A, K, \delta)$ by (30) for different forward operators, with the choice of $\delta = 15$ which is sufficient for the lower bound to hold with probability at least 0.9 for all these forward operators. We present the result in Figure 6(b) and compare it with the SA factors presented in Figure 6(a). We find that our theoretically justified lower bound is still able to distinguish well whether a given inverse problem is suitable or not for stochastic optimization, but seems to be very conservative for the choice of δ. The result in Figure 6(c) suggests that if a smaller δ is chosen heuristically, one may scale up the result in Figure 6(b) and obtain a better lower bound estimate for the SA factors. We conjecture that the probabilistic result in Theorem V.3 may be further improved – this is an open question for future work.

For a better demonstration, we additionally provide experiments on two inherently different imaging inverse problems: the space-varying deblurring and X-ray CT image reconstruction. We will see that as suggested by the distinct SA factors for these two inverse problems, numerically the acceleration provided by stochastic gradient techniques indeed varies in these two problems for a wide range of algorithms. Moreover, for some inverse problems such as deblurring, the minibatch partition scheme we use will play an important role in the actual convergence rate of the stochastic gradient methods, as suggested by our theory. We use here a classic form of smooth edge-preserving regularization [31], [76] which is widely used for these inverse problems:

$$x^\star \in \arg\min_{x \in \mathcal{K}} \left\{ \frac{1}{2n} \|Ax - b\|_2^2 + \lambda \sum_r \phi([Dx]_r) \right\}, \quad (33)$$

where D is the 2D differential operator, and $\phi(\cdot)$ is an edge-preserving potential function which penalize the differences between neighboring pixels [31], [76]–[78], while the constraint set $\mathcal{K}$ here is the box constraint restricting the resulting image to have pixel values between 0 and 1. We choose to use the potential function proposed by Thibault et al [76]:

$$\phi(z) = \frac{|z|^p}{1 + |\frac{z}{c}|^{p-q}}, \quad (34)$$

which encompasses the well-known approximate Huber prior ($p = 2, q = 1$) and the generalized Gaussian markov field ($1 < q = p \leq 2$). Here we set $p = 2, q = 1.5, c = 10$. We test the performance of the Katyusha algorithm [21] which is an accelerated stochastic variance-reduced gradient method, and the modified FISTA of Chambolle and Dossal [10] with provable convergence on iterates. Since the Katyusha algorithm uses variance-reduced stochastic gradients, its step-size is non-vanishing and hence converge faster than the non-variance-reduced stochastic methods in high-accuracy regimes [39]. For all the compared algorithms, we use the step-sizes suggested by their theoretical convergence analysis. We also partition the smooth regularizer into the minibatches for the Katyusha algorithm, such that the computational costs for 1 datapass (epoch) of Katyusha and FISTA are equivalent in our examples.

We first test the algorithms on a space-varying deblurring task for images sized 512 by 512, with a space-varying out-of-focus blur kernel. As suggested by our lower bound $\alpha_\ell(A, \bar{\mathcal{I}}, K)$ in Theorem V.2 and the numerical result shown in Figure 4, we choose the checkerboard subsampling minibatch partition with $K = 20$ for the deblurring task which has small local-accumulated-coherence for Katyusha and compare it with the random partition scheme. Moreover, we are aware of that, due to the presence of (restricted) strong-convexity, the convergence for both of the accelerated deterministic and stochastic methods can be further improved by restart schemes in a local high-accuracy regime [79]. Hence we also compare the restart variants of Katyusha [80] and FISTA [79] where we make grid search for the best restart period for each algorithm. All algorithms are initialized with a backprojection.

We plot the root mean square distance (RMSD) towards the minimizer $\frac{\|x - x^\star\|_2}{\sqrt{d}}$ in Figure 7 for each algorithm, where

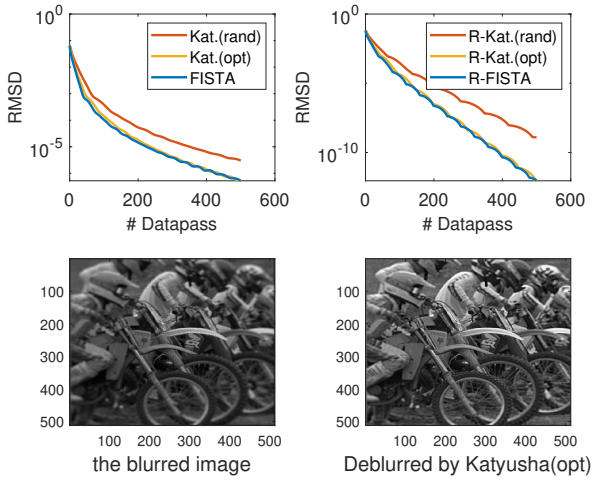


Fig. 7: *The root mean square distance plot for the deblurring experiment with a smooth edge-preserving regularization. Image: Kodim05, with an additive Guassian noise (variance 1).*

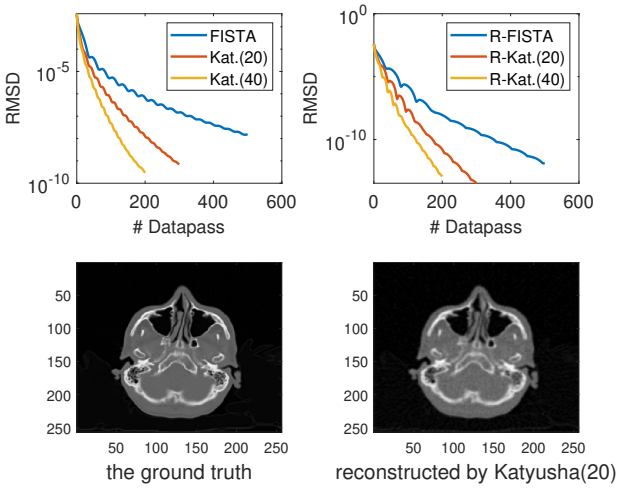


Fig. 8: *The root mean square distance plot for the X-ray CT image reconstruction experiment with a smooth edge-preserving regularization.* $\log_{10} \frac{\|Ax^\dagger\|_2^2}{\|w\|_2^2} \approx 3.16$.

$x^\star$ denotes the minimizer of the objective function (33)[6]. We can clearly observe that the checkerboard interleaving minibatch scheme leads to a superior performance over the random minibatch for Katyusha, as predicted by the theory. We also report that in this experiment, if we further increase the number of minibatches of Katyusha and Restart-Katyusha, we do not observe faster convergence. In this experiment, no matter how we increase the number of subsets, we do not observe improved performance of Katyusha over FISTA – such a trend is successfully predicted by the SA factor shown in the Figure 3, where we can see that the curve of the SA factor for deblurring task has low values, and goes flat instead of increasing after the number of minibatches $K > 10$.

We also consider a 2D fan-beam CT imaging problem generated via the Matlab package *AIRtools* [81], where we aim to reconstruct a 256×256 head image from 92532 noisy X-ray measurements (hence the forward operator $A \in \mathbb{R}^{92532 \times 65536}$), using the smooth edge-preserving regularization. Denoting $x^\dagger$ to be the (vectorized) ground truth image and $w \in \mathbb{R}^n$ to be an additional random noise vector drawn from an exponential Poisson distribution, we have the observed measurement as $b = Ax^\dagger + w$. The signal-to-noise ratio of the X-ray measurement in this example is set to be: $\log_{10} \frac{\|Ax^\dagger\|_2^2}{\|w\|_2^2} \approx 3.16$. For the stochastic methods, we use a partition scheme based on interleaving views from equally-spaced angle (20) which is standard and most practical for CT reconstruction [29].

We present the convergence results of the compared algorithms in Figure 8. Unlike in the previous deblurring example, in this experiment we observe that Katyusha/Restart-Katyusha are significantly faster than the deterministic methods, especially when we use a relatively large number of minibatches ($K = 40$), due to the fact that CT imaging has a large SA factor, as our analysis suggested. The comparative results of Restart-Katyusha and Restart-FISTA in deblurring and CT suggest that our proposed SA factor is also useful in the case where the (restricted) strong-convexity condition holds and is exploited by the algorithms, and we refer the readers to the supplemental material where we include an analysis regarding this case.

[6]Here we choose to use the RMSD to the minimizer instead of the mean-square-error to the ground-truth – the RMSD is more appropriate because the primary focus of this work is on optimization instead of estimation.

VII. Conclusion

In this work we have investigated the practicability of the state-of-the-art stochastic gradient methods in imaging inverse problems. We first presented a negative result on existing SGD-type methods on image deblurring, as a motivational example. To understand the limitation of stochastic gradient methods in inverse problems, we analyzed the worse-case theoretical limits and proposed the SA factor to evaluate the possible computational advantage of using stochastic techniques for a given task. Then we found that the SA factor is directly related to the inherent structure of the forward measurement model.

We derived lower and upper bounds of the SA factor. From the theoretical results, we found out that, if a linear inverse problem has a small ratio of $\frac{\|A^T\|_{1\to 2}^2}{\|A\|^2}$, which means the Hessian matrix $A^T A$ has fast-decaying eigenspectrum, then it typically admits good SA factors, hence can be rapidly solved by stochastic gradient methods. Our analysis also suggests that, excellent partition schemes typically have low local-accumulated-coherence, which essentially means the measurements within one minibatch have low mutual correlation. Using the SA factor, jointly with the derived lower bounds, practitioners can identify whether they should use stochastic gradient or deterministic gradient algorithms for given inverse problems, and evaluate the potential of given partition schemes. Our result also provides intuition that minibatch partitions with low coherence are superior than the ones with

higher coherence, suggesting that practitioners may use this as a criterion to design heuristic schemes for generating near-optimal partitions (for example the checkerboard partitioning for deblurring and the view-based interleaving for CT). However, exactly finding the best partitioning for a given inverse problem is a combinatorial problem, and we do not currently have any efficient generic scheme which is guaranteed to exactly solve this for arbitrary inverse problems – we leave this as an open problem for future work.

It is worth noting a limitation of our work. Our bounds do not take into account the fact that, for some inverse problems such as PET and CT, it is possible to judiciously design an ordering for accessing the minibatches, which can be empirically superior to the random access ordering [28], [29]. The empirical benefits of ordered-subsets methods over stochastic gradient methods in these applications have not yet been theoretically understood – this suggests an important and interesting line of future work might be to study SGD techniques where tight theoretical convergence rate analysis is possible, along with non-iid sampling strategies (e.g. using Markov chain sampling of the mini-batches [82]), in imaging inverse problems.

While our results are mainly for linear inverse problems with least-squares data-fidelity terms, we believe that they also can be extended and give insights to inverse problems with non-linear measurements since one can construct majorizing linearized subproblems (proximal Newton-steps) and solve these subproblems with deterministic or stochastic proximal gradient methods. Our results may also be extended for understanding and analyzing the limitations of stochastic gradient-based methods [52] with the plug-and-play priors [50], [51] and the regularization-by-denoising schemes [53]–[55] in imaging inverse problems, which we leave as a future direction.

Although we have concentrated on stochastic gradient methods vs deterministic gradient methods, there are other considerations that might affect the choice of whether to go stochastic. For example, if an inverse problem can be effectively preconditioned by simple preconditioners (such as diagonal preconditioners) or implicitly by variable-metric optimization techniques, then the potential benefit of stochastic methods over deterministic methods may possibly be reduced, since the preconditioned forward operator may not have as fast-decaying spectrum as the original one. Moreover, if the forward operator can be implemented with a fast transform such as the FFT, for example in MRI image reconstruction tasks, the deterministic gradient methods are usually much more favored since they can benefit from the fast operation while current stochastic gradient methods cannot. We consider these as topics for future research.

APPENDIX A
THE PROOF OF THEOREM IV.3

We set $x^0 \in \mathcal{X}$ as initialization for both $\mathcal{A}_{\text{full}}$ and $\mathcal{A}_{\text{stoc}}$, where $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2^2 \leq 1\right\}$. According to the lower bound for the stochastic gradient by [68, Theorem 7], there exists an objective function $F(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x)$ with a set of convex and L_b-smooth function f_i, for a positive constant C_{stoc}, which is independent of L_b, $L_{\nabla f}$ and K, such that in order to achieve an output $\mathbb{E}F(x^t_{\mathcal{A}}) - F(x^\star) \leq \epsilon$, any stochastic gradient algorithm must take at least $C_{\text{stoc}}\left(K + \sqrt{\frac{KL_b}{\epsilon}}\right)$ calls of the stochastic gradient oracle $\nabla f_i()$. In other words, for this worst case function, if we run any stochastic gradient method with only Kt calls on the stochastic gradient oracle such that $Kt = C_{\text{stoc}}\sqrt{\frac{KL_b}{\epsilon}}$, then $\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star) \geq \epsilon$ can be guaranteed. Hence, we have:

$$\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star) \geq \frac{C^2_{\text{stoc}}L_b}{Kt^2} \tag{35}$$

Meanwhile, starting from $x^0 \in \mathcal{X}$, by Def. IV.1, for any optimal full gradient method $\mathcal{A}_{\text{full}}$ we can have:

$$F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star) \leq \frac{C_1 L_{\nabla f}\|x^0 - x^\star\|_2^2}{t^2} \leq \frac{4C_1 L_{\nabla f}}{t^2}, \tag{36}$$

where the constant C_1 is independent of L_b, $L_{\nabla f}$ and K. Combining these two bounds we can have:

$$\frac{\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)}{F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star)} \geq \frac{C^2_{\text{stoc}}L_b}{4C_1 K L_{\nabla f}}. \tag{37}$$

Finally, by setting $c_0 = \frac{C^2_{\text{stoc}}}{4C_1}$ we yield the lower bound.

Next, we are going to prove the upper bound. According to the lower bound for the deterministic gradient by [68, Theorem 3], there exists an objective function $F(x) = \frac{1}{K}\sum_{k=1}^{K} f_k(x)$ with a set of convex and L_b-smooth function f_i, such that in order to achieve an output $F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star) \leq \epsilon$, any deterministic gradient algorithm must take at least $\Omega\left(\sqrt{\frac{L_{\nabla f}}{\epsilon}}\right)$ calls of the deterministic gradient oracle $\nabla f(\cdot)$. Hence there exists a positive constant C_{full}, which is independent of L_b, $L_{\nabla f}$ and K, such that if we take $t = C_{\text{full}}\sqrt{\frac{L_{\nabla f}}{\epsilon}}$ number of calls on deterministic gradient oracle, we are guaranteed to have $F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star) \geq \epsilon$. Then we have:

$$F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star) \geq \frac{C^2_{\text{full}}L_{\nabla f}}{t^2} \geq \frac{C^2_{\text{full}}L_{\nabla f}\|x^0 - x^\star\|_2^2}{t^2} \tag{38}$$

Meanwhile, for optimal stochastic algorithm we have the upper bound of convergence by Def. IV.2 with setting $m = K$:

$$\begin{aligned}&\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)\\ &\leq \frac{C_2(F(x^0) - F(x^\star)) + \frac{C_3 L_b}{K}\|x^0 - x^\star\|_2^2}{t^2}\end{aligned} \tag{39}$$

Combining the two bounds we can have:

$$\frac{\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)}{F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star)} \leq \frac{C_2}{C^2_{\text{full}}}\cdot\frac{L_b}{KL_{\nabla f}} + \frac{C_3(F(x^0) - F(x^\star))}{C^2_{\text{full}}L_{\nabla f}\|x^0 - x^\star\|_2^2}. \tag{40}$$

Recall the definition of smoothness, we can have:

$$f(x^0) - f(x^\star) - \langle \nabla f(x^\star), x^0 - x^\star\rangle \leq \frac{L_{\nabla f}}{2}\|x^0 - x^\star\|_2^2. \tag{41}$$

Since the solution set of worst-case objective function proposed in [68, Theorem 3] lives in the relative interior of $\mathcal{X}$,

we can have $\triangledown f(x^\star) = 0$, and hence $F(x^0) - F(x^\star) = f(x^0) - f(x^\star) \leq \frac{L_{\triangledown f}}{2}\|x^0 - x^\star\|_2^2$. Consequently, we can have:

$$\frac{\mathbb{E}F(x^t_{\mathcal{A}_{\text{stoc}}}) - F(x^\star)}{F(x^t_{\mathcal{A}_{\text{full}}}) - F(x^\star)} \leq \frac{C_2}{C_{\text{full}}^2} \cdot \frac{L_b}{KL_{\triangledown f}} + \frac{C_3}{2C_{\text{full}}^2}. \tag{42}$$

By setting $c_1 = \frac{C_2}{C_{\text{full}}^2}$ and $c_2 = \frac{C_3}{2C_{\text{full}}^2}$ we yield the upper bound.

Appendix B
The proof of Theorem V.2

If we set $H = S^k A(S^k A)^T$ for some $k \in [K]$, the top eigenvalue of $S^k A(S^k A)^T$ is no larger than the largest value within the set $G(S^k A(S^k A)^T)$ which we denote here as $G_{\max}(S^k A(S^k A)^T)$. We have the following relationship:

$$\begin{aligned}\|S^k A\|^2 \leq G_{\max}(S^k A(S^k A)^T) &= \max_{i\in\mathcal{I}_k} \|(S^k A)a_i\|_1 \\ &= \max_{i\in\mathcal{I}_k} \sum_{j\in\mathcal{I}_k} |\langle a_i, a_j\rangle|.\end{aligned} \tag{43}$$

Then we have:

$$L_b = \frac{K}{n} \max_{k\in[K]} \|S^k A(S^k A)^T\| \leq \frac{K}{n} \max_{q\in[K]} \max_{i\in\mathcal{I}_q} \sum_{j\in\mathcal{I}_q} |\langle a_i, a_j\rangle|, \tag{44}$$

hence $L_b \leq \frac{K}{n}\mu_\ell(A, \bar{\mathcal{I}}, K)$. By definition of the SA factor, we can write:

$$\Upsilon(A, \bar{\mathcal{I}}, K) = \frac{KL_{\triangledown f}}{L_b} \geq \frac{\|A\|^2}{\mu_\ell(A, \bar{\mathcal{I}}, K)}. \tag{45}$$

On the other hand, note that by the definition of the local accumulated coherence, we can have an upper bound for $\mu_\ell(A, \bar{\mathcal{I}}, K) := \max_{q\in[K]} \max_{i\in\mathcal{I}_q} \sum_{j\in\mathcal{I}_q} |\langle a_i, a_j\rangle|$:

$$\mu_\ell(A, \bar{\mathcal{I}}, K) \leq \frac{n}{K} \max_{i\in[n]} \|a_i\|_2^2 = \frac{n}{K}\|A^T\|_{1\to 2}^2, \tag{46}$$

and hence we can have a relaxed lower bound for $\Upsilon(A, \bar{\mathcal{I}}, K)$:

$$\Upsilon(A, \bar{\mathcal{I}}, K) \geq \frac{\|A\|^2}{\mu_\ell(A, \bar{\mathcal{I}}, K)} \geq \frac{K\|A\|^2}{n\|A^T\|_{1\to 2}^2}. \tag{47}$$

Suppose that for some positive constant ρ we have:

$$\rho := \frac{\max_{i\in[n]} \|a_i\|_2^2}{\frac{1}{n}\sum_{j=1}^n \|a_j\|_2^2}, \tag{48}$$

then we can write:

$$\max_{i\in[n]} \|a_i\|_2^2 \leq \frac{\rho}{n} \sum_{j=1}^n \|a_j\|_2^2 = \frac{\rho\|A\|_F^2}{n} = \frac{\rho\sum_{i=1}^d \sigma_i(A^T A)}{n}, \tag{49}$$

and hence we can further lower bound $\Upsilon(A, \bar{\mathcal{I}}, K)$ by the cumulative eigenspectrum of the Hessian:

$$\Upsilon(A, \bar{\mathcal{I}}, K) \geq \frac{K\|A\|^2}{n\|A^T\|_{1\to 2}^2} \geq \frac{K \cdot \sigma_1(A^T A)}{\rho \cdot \sum_{i=1}^d \sigma_i(A^T A)}. \tag{50}$$

thus finishes the proof.

Appendix C
The proof of Theorem V.4

[73, Theorem 4.3.15] indicates that, for a given Hermitian matrix $H \in \mathbb{R}^{n\times n}$, and any of its m-by-m principal submatrices H_m, obtained by deleting $n - m$ rows and columns from H, we can have:

$$\sigma_1(H_m) \geq \sigma_{n-m+1}(H). \tag{51}$$

If we set $H_m = S^k A(S^k A)^T$, then we have:

$$\|S^k A(S^k A)^T\| = \|S^k (AA^T) S^{k^T}\| \geq \sigma_{n-m+1}(AA^T). \tag{52}$$

Now we use the fact that $S^k A(S^k A)^T$ and $(S^k A)^T S^k A$ share the same non-zero eigenvalues, and meanwhile AA^T and $A^T A$ also shares the same non-zero eigenvalues, we can have the following bound:

$$\begin{aligned}\|(S^k A)^T S^k A\| = \|S^k A(S^k A)^T\| &\geq \sigma_{n-m+1}(AA^T) \\ &= \sigma_{n-m+1}(A^T A).\end{aligned} \tag{53}$$

Then by the definition of $\Upsilon(A, \bar{\mathcal{I}}, K)$ we can obtain the upper bound.

Appendix D
The proof of Theorem V.3

Suppose we randomly permute the index $[n]$ and generate the partition index $[\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_K]$. If we pick arbitrarily a number $k \in [K]$ where S^k is the subsampling matrix, by the Matrix Chernoff inequality [75] we have $\|A^T S^{k^T} S^k A\| \leq (1 + \delta_0) \cdot \frac{\|A\|^2}{K}$ with probability at least:

$$P := 1 - d \cdot \left[\frac{e^{\delta_0}}{(\delta_0 + 1)^{\delta_0 + 1}}\right]^{\frac{\|A\|^2}{K\|A^T\|_{1\to 2}^2}}, \tag{54}$$

for any $\delta_0 > 0$. Now by choosing $\delta_0 = \delta \cdot \frac{K\|A^T\|_{1\to 2}}{\|A\|^2}$ we can have following:

$$\|A^T S^{k^T} S^k A\| \leq (1 + \delta \cdot \frac{K\|A^T\|_{1\to 2}}{\|A\|^2}) \cdot \frac{\|A\|^2}{K} = \frac{\|A\|^2}{K} + \delta\|A^T\|_{1\to 2}^2, \tag{55}$$

with probability at least P' where:

$$1 - d \cdot \left[\frac{e^\delta}{(\delta \cdot \frac{K\|A^T\|_{1\to 2}}{\|A\|^2})^\delta}\right] \geq P' := 1 - d \cdot \left[\frac{e^\delta}{\delta^\delta}\right]. \tag{56}$$

(This is because we restrict here $K \geq \frac{\|A\|^2}{\|A^T\|_{1\to 2}^2}$.) Now by applying the union bound over all possible choices of k and since we assume here $K \leq \min(n, d)$, we have, with probability at least $1 - d^2 \cdot \left[\frac{e^\delta}{\delta^\delta}\right]$:

$$\max_{k\in[K]} \|A^T S^{k^T} S^k A\| \leq \frac{\|A\|^2}{K} + \delta\|A^T\|_{1\to 2}^2. \tag{57}$$

Then by definition $\Upsilon(A, \bar{\mathcal{I}}, K) = \frac{KL_{\triangledown f}}{L_b}$ and hence:

$$\Upsilon(A, \bar{\mathcal{I}}, K) \geq \frac{\frac{K}{n}\|A\|^2}{\frac{K}{n}\left(\frac{\|A\|^2}{K} + \delta\|A^T\|_{1\to 2}^2\right)} = \frac{1}{\frac{1}{K} + \delta\frac{\|A^T\|_{1\to 2}^2}{\|A\|^2}} \tag{58}$$

Now since

$$\|A^T\|_{1\to 2}^2 = \max_{i\in[n]} \|a_i\|_2^2 \leq \frac{\rho}{n}\sum_{j=1}^{n}\|a_j\|_2^2 = \frac{\rho\sum_{i=1}^{d}\sigma_i(A^TA)}{n}, \tag{59}$$

we have:

$$\Upsilon(A,\bar{\mathcal{I}},K) \geq \frac{1}{\frac{1}{K}+\delta\frac{\|A^T\|_{1\to 2}^2}{\|A\|^2}} \geq \frac{1}{\frac{1}{K}+\delta\rho\cdot\frac{\frac{\sum_{i=1}^{d}\sigma_i(A^TA)}{n}}{\sigma_1(A^TA)}}. \tag{60}$$

Thus finishes the proof.

REFERENCES

[1] J. Tang, K. Egiazarian, and M. Davies, "The limitation and practical acceleration of stochastic gradient algorithms in inverse problems," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7680–7684.

[2] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[3] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proceedings of 3rd International Conference on Learning Representations*, 2015.

[5] A. Chambolle and T. Pock, "An introduction to continuous optimization for imaging," *Acta Numerica*, vol. 25, pp. 161–319, 2016.

[6] P.-L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

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

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

[9] ——, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2419–2434, 2009.

[10] A. Chambolle and C. Dossal, "On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"," *Journal of Optimization theory and Applications*, vol. 166, no. 3, pp. 968–982, 2015.

[11] J. Liang and C.-B. Schönlieb, "Improving FISTA: Faster, smarter and greedier," *arXiv preprint arXiv:1811.01430*, 2018.

[12] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of mathematical imaging and vision*, vol. 40, no. 1, pp. 120–145, 2011.

[13] ——, "On the ergodic convergence rates of a first-order primal–dual algorithm," *Mathematical Programming*, vol. 159, no. 1-2, pp. 253–287, 2016.

[14] B. Jin, Z. Zhou, and J. Zou, "On the convergence of stochastic gradient descent for nonlinear ill-posed problems," *SIAM Journal on Optimization*, vol. 30, no. 2, pp. 1421–1450, 2020.

[15] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, pp. 1–30, 2013.

[16] 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.

[17] L. Xiao and T. Zhang, "A proximal stochastic gradient method with progressive variance reduction," *SIAM Journal on Optimization*, vol. 24, no. 4, pp. 2057–2075, 2014.

[18] S. Shalev-Shwartz and T. Zhang, "Stochastic dual coordinate ascent methods for regularized loss minimization," *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 567–599, 2013.

[19] Y. Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," in *Soviet Mathematics Doklady*, vol. 27, no. 2, 1983, pp. 372–376.

[20] G. Lan and Y. Zhou, "An optimal randomized incremental gradient method," *Mathematical programming*, vol. 171, no. 1-2, pp. 167–215, 2018.

[21] Z. Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 8194–8244, 2017.

[22] K. Zhou, F. Shang, and J. Cheng, "A simple stochastic variance reduced algorithm with fast convergence rates," in *International Conference on Machine Learning*, 2018, pp. 5975–5984.

[23] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 2783–2808, 2018.

[24] M. J. Ehrhardt, P. Markiewicz, and C.-B. Schönlieb, "Faster pet reconstruction with non-smooth priors by randomization and preconditioning," *Physics in Medicine & Biology*, vol. 64, no. 22, p. 225019, 2019.

[25] D. Karimi and R. K. Ward, "A hybrid stochastic-deterministic gradient descent algorithm for image reconstruction in cone-beam computed tomography," *Biomedical Physics & Engineering Express*, vol. 2, no. 1, p. 015008, 2016.

[26] ——, "Sparse-view image reconstruction in cone-beam computed tomography with variance-reduced stochastic gradient descent and locally-adaptive proximal operation," *Journal of Medical and Biological Engineering*, vol. 37, no. 3, pp. 420–440, 2017.

[27] H. Erdogan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Physics in Medicine & Biology*, vol. 44, no. 11, p. 2835, 1999.

[28] G. T. Herman and L. B. Meyer, "Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application)," *IEEE transactions on medical imaging*, vol. 12, no. 3, pp. 600–609, 1993.

[29] D. Kim, S. Ramani, and J. A. Fessler, "Combining ordered subsets and momentum for accelerated x-ray ct image reconstruction," *IEEE transactions on medical imaging*, vol. 34, no. 1, pp. 167–178, 2015.

[30] M. Muckley, D. C. Noll, and J. A. Fessler, "Accelerating sense-type mr image reconstruction algorithms with incremental gradients," in *Proc. Intl. Soc. Mag. Res. Med*, 2014, p. 4400.

[31] S. Sotthivirat and J. A. Fessler, "Relaxed ordered-subset algorithm for penalized-likelihood image restoration," *JOSA A*, vol. 20, no. 3, pp. 439–449, 2003.

[32] J. Douglas and H. H. Rachford, "On the numerical solution of heat conduction problems in two and three space variables," *Transactions of the American mathematical Society*, vol. 82, no. 2, pp. 421–439, 1956.

[33] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[34] P. L. Combettes and J.-C. Pesquet, "Stochastic quasi-fejér block-coordinate fixed point iterations with random sweeping," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 1221–1248, 2015.

[35] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[36] S. Becker and J. Fadili, "A quasi-newton proximal splitting method," in *Advances in Neural Information Processing Systems*, 2012, pp. 2618–2626.

[37] E. Chouzenoux, J.-C. Pesquet, and A. Repetti, "Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function," *Journal of Optimization Theory and Applications*, vol. 162, no. 1, pp. 107–132, 2014.

[38] E. Chouzenoux and J.-C. Pesquet, "A stochastic majorize-minimize subspace algorithm for online penalized least squares estimation," *IEEE Transactions on Signal Processing*, vol. 65, no. 18, pp. 4770–4783, 2017.

[39] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in neural information processing systems*, 2013, pp. 315–323.

[40] R. M. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik, "SGD: General analysis and improved rates," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. Long Beach, California, USA: PMLR, 09–15 Jun 2019, pp. 5200–5209.

[41] R. Sun, "Optimization for deep learning: theory and algorithms," *arXiv preprint arXiv:1912.08957*, 2019.

[42] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[43] H. Erdogan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Physics in Medicine and Biology*, vol. 44, no. 11, p. 2835, 1999. [Online]. Available: http://stacks.iop.org/0031-9155/44/i=11/a=311

[44] K. Wei, R. K. Iyer, S. Wang, W. Bai, and J. A. Bilmes, "Mixed robust/average submodular partitioning: Fast algorithms, guarantees, and applications," in *Advances in Neural Information Processing Systems*, 2015, pp. 2233–2241.

[45] S. Wang, W. Bai, C. Lavania, and J. Bilmes, "Fixing mini-batch sequences with hierarchical robust partitioning," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 3352–3361.

[46] P. Zhao and T. Zhang, "Stochastic optimization with importance sampling for regularized loss minimization," in *international conference on machine learning*, 2015, pp. 1–9.

[47] Z. Allen-Zhu, Z. Qu, P. Richtárik, and Y. Yuan, "Even faster accelerated coordinate descent using non-uniform sampling," in *International Conference on Machine Learning*, 2016, pp. 1110–1119.

[48] D. Csiba and P. Richtárik, "Importance sampling for minibatches," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 962–982, 2018.

[49] K. Egiazarian, A. Foi, and V. Katkovnik, "Compressed sensing image reconstruction via recursive spatially adaptive filtering," in *2007 IEEE International Conference on Image Processing*, vol. 1. IEEE, 2007, pp. I–549.

[50] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 2013, pp. 945–948.

[51] U. S. Kamilov, H. Mansour, and B. Wohlberg, "A plug-and-play priors approach for solving nonlinear imaging inverse problems," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1872–1876, 2017.

[52] Y. Sun, B. Wohlberg, and U. S. Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Transactions on Computational Imaging*, 2019.

[53] Y. Romano, M. Elad, and P. Milanfar, "The little engine that could: Regularization by denoising (red)," *SIAM Journal on Imaging Sciences*, vol. 10, no. 4, pp. 1804–1844, 2017.

[54] E. T. Reehorst and P. Schniter, "Regularization by denoising: Clarifications and new interpretations," *IEEE Transactions on Computational Imaging*, vol. 5, no. 1, pp. 52–67, 2018.

[55] Z. Wu, Y. Sun, A. Matlock, J. Liu, L. Tian, and U. S. Kamilov, "Simba: scalable inversion in optical tomography using deep denoising priors," *IEEE Journal of Selected Topics in Signal Processing*, 2020.

[56] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image restoration by sparse 3d transform-domain collaborative filtering," in *Image Processing: Algorithms and Systems VI*, vol. 6812. International Society for Optics and Photonics, 2008, p. 681207.

[57] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017.

[58] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[59] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "Fast image recovery using variable splitting and constrained optimization," *IEEE transactions on image processing*, vol. 19, no. 9, pp. 2345–2356, 2010.

[60] O. Whyte, J. Sivic, A. Zisserman, and J. Ponce, "Non-uniform deblurring for shaken images," *International journal of computer vision*, vol. 98, no. 2, pp. 168–186, 2012.

[61] Kodak-Lossless-True-Color-Image-Suite. [Online]. Available: http://r0k.us/graphics/kodak/

[62] M. S. Almeida and M. Figueiredo, "Deconvolving images with unknown boundaries using the alternating direction method of multipliers," *IEEE Transactions on Image processing*, vol. 22, no. 8, pp. 3074–3086, 2013.

[63] N. Perraudin, V. Kalofolias, D. Shuman, and P. Vandergheynst, "Unlocbox: A matlab convex optimization toolbox for proximal-splitting methods," *arXiv preprint arXiv:1402.0779*, 2014.

[64] L. Rosasco, S. Villa, and B. C. Vũ, "Convergence of stochastic proximal gradient algorithm," *Applied Mathematics & Optimization*, pp. 1–27, 2019.

[65] J. Konečnỳ, J. Liu, P. Richtárik, and M. Takáč, "Mini-batch semi-stochastic gradient descent in the proximal setting," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 242–255, 2016.

[66] Y. Nesterov, "Gradient methods for minimizing composite objective function," UCL, Tech. Rep., 2007.

[67] J. M. Fadili and G. Peyré, "Total variation projection with first order schemes," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 657–669, 2010.

[68] B. E. Woodworth and N. Srebro, "Tight complexity bounds for optimizing composite objectives," in *Advances in neural information processing systems*, 2016, pp. 3639–3647.

[69] D. Kim and J. A. Fessler, "On the convergence analysis of the optimized gradient method," *Journal of optimization theory and applications*, vol. 172, no. 1, pp. 187–205, 2017.

[70] A. B. Taylor, J. M. Hendrickx, and F. Glineur, "Exact worst-case performance of first-order methods for composite convex optimization," *SIAM Journal on Optimization*, vol. 27, no. 3, pp. 1283–1313, 2017.

[71] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Information theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[72] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE transactions on medical imaging*, vol. 13, no. 4, pp. 601–609, 1994.

[73] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge university press, 2012.

[74] D. Needell and J. A. Tropp, "Paved with good intentions: analysis of a randomized block kaczmarz method," *Linear Algebra and its Applications*, vol. 441, pp. 199–221, 2014.

[75] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of computational mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[76] J.-B. Thibault, K. D. Sauer, C. A. Bouman, and J. Hsieh, "A three-dimensional statistical approach to improved image quality for multislice helical ct," *Medical physics*, vol. 34, no. 11, pp. 4526–4544, 2007.

[77] A. H. Delaney and Y. Bresler, "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography," *IEEE Transactions on Image Processing*, vol. 7, no. 2, pp. 204–221, 1998.

[78] T. Pock and S. Sabach, "Inertial proximal alternating linearized minimization (ipalm) for nonconvex and nonsmooth problems," *SIAM Journal on Imaging Sciences*, vol. 9, no. 4, pp. 1756–1787, 2016.

[79] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Foundations of computational mathematics*, vol. 15, no. 3, pp. 715–732, 2015.

[80] J. Tang, M. Golbabaee, F. Bach, and M. E. davies, "Rest-katyusha: Exploiting the solution's structure via scheduled restart schemes," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 427–438.

[81] P. C. Hansen and M. Saxild-Hansen, "AIR tools—a MATLAB package of algebraic iterative reconstruction methods," *Journal of Computational and Applied Mathematics*, vol. 236, no. 8, pp. 2167–2178, 2012.

[82] T. Sun, Y. Sun, and W. Yin, "On markov chain gradient descent," in *Advances in Neural Information Processing Systems*, 2018, pp. 9896–9905.