THE LIMITATION AND PRACTICAL ACCELERATION OF STOCHASTIC GRADIENT ALGORITHMS IN INVERSE PROBLEMS

Junqi Tang[†], *Karen Egiazarian*[⋆∗], *Mike Davies*[†]

[†] School of Engineering, University of Edinburgh, UK; [⋆]Noiseless Imaging Ltd, Finland.
[∗]Faculty of Computing and Electrical Engineering, Tampere University, Finland

ABSTRACT

In this work we investigate the practicability of stochastic gradient descent and recently introduced variants with variance-reduction techniques in imaging inverse problems, such as space-varying image deblurring. Such algorithms have been shown in machine learning literature to have optimal complexities in theory, and provide great improvement empirically over the full gradient methods. Surprisingly, in some tasks such as image deblurring, many of such methods fail to converge faster than the accelerated full gradient method (FISTA), even in terms of epoch counts. We investigate this phenomenon and propose a theory-inspired mechanism to characterize whether a given inverse problem should be preferred to be solved by stochastic optimization technique with a known sampling pattern. Furthermore, to overcome another key bottleneck of stochastic optimization which is the heavy computation of proximal operators while maintaining fast convergence, we propose an accelerated primal-dual SGD algorithm and demonstrate the effectiveness of our approach in image deblurring experiments.

Index Terms— Stochastic Optimization, Inverse Problems, Image Processing

1. INTRODUCTION

The stochastic gradient methods [1, 2] and recently introduced variants with variance-reduction [3, 4, 5] have been widely used to solve large-scale convex optimization problem in machine learning applications. Such tasks can be formulated as the following:

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

where $\mathcal{X} \in \mathbb{R}^d$ is a close convex set and we denote $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$ the data fidelity term. Each $f_i(x)$ is assumed to be convex and L-smooth, while the regularization term $g(x)$ is a simple convex function and is possibly non-smooth. With Nesterov's acceleration [6, 7], researchers [8, 9, 10, 11] have developed several "optimal" algorithms which can provably achieve the worse-case optimal convergence rate for (1).

While having been a proven success both in theory and in machine learning applications, there is no convincing result so far in the literature which reports the performance of the stochastic gradient methods in image processing applications (except for tomography reconstruction [12, 13, 14]), which also involve large-scale optimization tasks in the same form of (1). In this work we investigate the practical performance of such methods, using space-varying deblurring as a running example.

We make the following contributions:

(Evaluating the limitation of stochastic gradient algorithms.) We investigate the fundamental limit of possible acceleration of a stochastic gradient method over its full gradient counterpart by measuring the *Stochastic Acceleration* (SA) factor which is based on the ratio of the Lipschitz constants of the minibatched stochastic gradient and the full gradient. We discover that the SA factor is indeed able to characterize the potential of a certain optimization task being speedily solved by applying randomization techniques.

(Breaking the computational bottleneck of expensive/multiple proximal operators for Nesterov-type momentum SGD.) Another factor in image processing practice which significantly affects the SGD-type methods' actual performance is the frequent calculation of the costly proximal operator for the regularization terms which have a linear operator, such as the TV semi-norm – SGD methods needs to calculate it much more frequently than full gradient methods. Moreover most of the fast SGD methods can not cope with more than one non-smooth regularization terms. To overcome these we propose an accelerated primal-dual SGD algorithm which can efficiently handle (1) regularization with a linear operator, (2) multiple regularization terms, while (3) maintaining Nesterov-type accelerated convergence speed in practice.

2. FAILURES OF STOCHASTIC OPTIMIZATION

We start by a simple space-varying deblurring [15] example where the central part (sized 128 by 128) of the "Kodim05" image from *Kodak Lossless True Color Image Suite* [16] is blurred with an space-varying blur kernel which imposes less blurring at the center but increasingly severe blurring towards the edge. 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, to get an estimation of the ground truth image. The algorithms we test in the experiments include the accelerated full gradient method FISTA [17], SGD with momentum [18], the proximal SVRG [19] and its accelerated variant, Katyusha [9].

Perhaps surprisingly, on this experiment we report a negative result for the randomized algorithms. The most efficient solver in this task is the full gradient method FISTA both in terms of wall clock time and number of datapasses. The state-of-the-art stochastic gradient method Katyusha even cannot beat FISTA in terms of epoch counts. For all the randomized algorithms we use a minibatch size which is 10 percent of the total data size. For stochastic gradient methods, smaller minibatch size in this case does not provide better performance in datapasses and will significantly slow down running time due to the multiple calls of proximal operator.

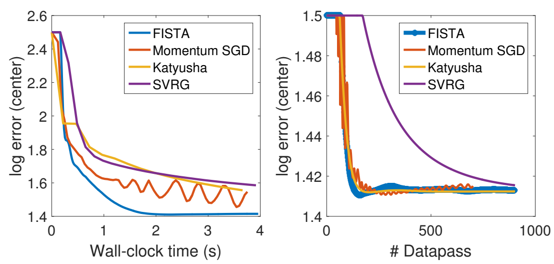

Fig. 1: The estimation error plot for the deblurring experiment. The plots correspond to the estimation error of the central part (100 by 100) of the image.

3. LIMITATIONS OF STOCHASTIC OPTIMIZATION

These results appear to be contrary to the popular belief among the stochastic optimization community, that stochastic gradient methods are much faster in terms of iteration complexity 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 $O\left(n + \sqrt{nL/\varepsilon}\right)$ evaluations of ∇f_i, while $O\left(n\sqrt{L/\varepsilon}\right)$ for optimal full gradient methods. Where is the loophole?

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

Given a minibatch index $[S_0, S_1, S_2, ..., S_K]$ such that:

$$f(x) = \frac{1}{K}\sum_{k=1}^{K} f_{S_k}(x), \quad f_{S_k}(x) := \frac{K}{n}\sum_{i \in S_k} f_i(x), \qquad (2)$$

In order to identify the potential of a certain optimization problem to be more efficiently solved by using stochastic gradient methods, we start by comparing the single iteration convergence of one instant of the proximal stochastic gradient descent with momentum named Katyusha [9] with the proximal accelerated full gradient descent (AFG), which read:

AFG(x_0, K, L)**:**

For $s = 0, 1, 2, ..., S$

$$x^s = \mathcal{T}_f(y^s, L) := \mathrm{prox}_{\lambda g}^{\frac{1}{L}}(y^s - \tfrac{1}{L}\nabla f(y^s));$$
$\rightarrow$ Proximal gradient descent
$$a_{s+1} = (1 + \sqrt{1 + 4a_s^2})/2;$$
$$y^{s+1} = x^s + \tfrac{a_s - 1}{a_{s+1}}(x^s - x^{s-1}); \rightarrow \text{Momentum}$$

Katyusha(x_0, S, m, L)**:**

For $s = 0, 1, 2, ..., S$

$$\theta = \tfrac{2}{s+4};$$
$$(\hat{x}^{s+1}, y^{s+1}, z^{s+1}) = \mathcal{A}(x^s, y^s, z^s, L, m, \theta, \nabla f(\hat{x}^s));$$

$\mathcal{A}(x^s, y^s, z^s, L, m, \theta, \nabla f(\hat{x}^s))$**:**

For $k = 0, 1, 2, ..., m$

$$x_{k+1} = \theta z_k + \tfrac{1}{2}\hat{x}^s + (\tfrac{1}{2} - \theta)y_k; \quad \rightarrow \text{Momentum}$$
Pick $i \in [1, 2, ...K]$ uniformly at random
$$\nabla_{k+1} = \nabla f(\hat{x}^s) + \nabla f_{S_i}(x_{k+1}) - \nabla f_{S_i}(\hat{x}^s);$$
$\rightarrow$ Compute a variance reduced stochastic gradient
$$z_{k+1} = \mathrm{prox}_{\lambda g}^{\frac{1}{3\theta L}}(z_k - \tfrac{1}{3\theta L}\nabla_{k+1});$$
$\rightarrow$ Proximal mirror descent
$$y_{k+1} = \mathrm{prox}_{\lambda g}^{\frac{1}{3L}}(x_{k+1} - \tfrac{1}{3L}\nabla_{k+1});$$
$\rightarrow$ Proximal gradient descent

where we define the proximal operator as:

$$\mathrm{prox}_{\lambda g}^{\eta}(\cdot) = \arg\min_{x \in \mathcal{X}} \frac{1}{2\eta}\|x - \cdot\|_2^2 + \lambda g(x). \qquad (3)$$

3.1. Analysis

We start with the standard smoothness assumption [20]:

A. 1 *(Smoothness of the Full-Batch and the Mini-Batches.) $f(.)$ is L_f-smooth and each f_{S_k} is L_m-smooth, that is:*

$$f(x) - f(y) - \nabla f(y)^T(x - y) \leq \frac{L_f}{2}\|x - y\|_2^2, \quad \forall a, b \in \mathcal{X}, \qquad (4)$$

and

$$f_{S_k}(x) - f_{S_k}(y) - \nabla f_{S_k}(y)^T(x - y) \leq \frac{L_b}{2}\|x - y\|_2^2, \qquad (5)$$

$\forall x, y \in \mathcal{X}$.

Now we are ready to present the main theorem, which follows from simply combining the existing convergence results of Katyusha and AFG, as well as the lower bounds for the stochastic and deterministic first-order optimization [20, 21].

Theorem 3.1 *Under **A.1**, let* $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\|_2^2, \leq 1\right\}$, $g(.) = 0$, $x_f^s = \mathcal{T}_f(y_f^{s-1}, L_f)$, $x_\mathcal{A}^s = \mathcal{A}(x_\mathcal{A}^{s-1}, L_b, m)$, $m = 2K$, $0 < s \leq \frac{d-1}{6}$, *there exist a set of convex and smooth function* $f_i(.)$, *such that* $\sum_{i=1}^K f_i = f \in \mathcal{F}_{L_f}^{\infty,1}(\mathbb{R}^d)$, *we have:*

$$\frac{\mathbb{E}f(x_\mathcal{A}^s) - f^\star}{f(x_f^{3s}) - f^\star} \leq 192 \cdot \frac{L_b}{KL_f} + \frac{512(f(x^0) - f^\star)}{L_f\|x^0 - x^\star\|_2^2} \tag{6}$$

and moreover, , there exist a set of convex and smooth function $f_i \in \mathcal{F}_{L_b}^{\infty,1}(\mathbb{R}^d)$, *such that with* $f = \sum_{i=1}^K f_i$ *we have,*

$$\frac{\mathbb{E}f(x_\mathcal{A}^s) - f^\star}{f(x_f^{3s}) - f^\star} \geq \frac{1}{11440} \cdot \frac{L_b}{KL_f}. \tag{7}$$

for a sufficiently large dimension $d = O(\frac{L_b^2}{\epsilon^2} \log \frac{L_b}{\epsilon} + K \log K)$ *where* $\mathbb{E}f(x_\mathcal{A}^s) - f^\star \leq \epsilon$.

From this theorem we can see that with the same epoch count (the iteration complexity of Katyusha's 1 epoch is equivalent to 3 iterations of AFG), the ratio of objective gap achieved by each algorithm can be upper and lower bounded by $\Theta(\frac{L_b}{KL_f})$ at the worst case. Although the constants seem pessimistic, it is within our expectation since the lower bounds on the convergence speed of both algorithms are derived on the worst possible function which satisfies **A.1**. Motivated by the theory, we propose to evaluate the potential of stochastic acceleration simply by the ratio $\frac{L_b}{KL_f}$ which dominates our upper and lower bounds in Theorem 3.1.

3.2. Evaluating the Limitation of SGD-type Algorithms

We introduced a metric called *Stochastic Acceleration* (SA) factor. The curve of SA factor as a function of the minibatch number K (for a given minibatch pattern) is able to provide a way of evaluating and characterizing inherently whether for a given inverse problem and a certain minibatch sampling scheme, randomized gradient methods should be preferred over the deterministic full gradient methods or not.

Definition 3.2 *For a given data-partitioning index* $\bar{S} = [S_1, ...S_K]$, *the Stochastic Acceleration (SA) factor is defined as:*

$$\Upsilon(\bar{S}) = \frac{KL_f}{L_b} \tag{8}$$

We test several least-squares loss $f(x) = \|Ax - b\|_2^2$ with different types of forward operator. In this case we have

$$f(x) = \|Ax - b\|_2^2 = \frac{1}{K}\sum_{k=1}^K f_{S_k}(x), \tag{9}$$

$$f_{S_k}(x) := K\|A_{S_k}x - b_{S_k}\|_2^2, \tag{10}$$

The examples of forward operator A we consider include the space-varying deblurring (262144 by 262144), a random compressed sensing matrix with i.i.d Guassian random entries (500 by 2000), a fan beam X-ray CT operator (91240 by 65536), and two machine learning datasets: RCV1 dataset (20242 by 47236), and Magic04 (19000 by 50, with random features). The data-partition we choose is the interleaving sampling. From the result show by the Fig 2 we find that indeed the stochastic methods have a limitation on some optimization problems like deblurring and inverse problems with random matrices, where we see that the curve of 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 the use of the ordered-subset methods [22]. The curve of SA factor on the left figure qualitatively predict the empirical comparison result of Katyusha and FISTA algorithm shown on the right, where we observe that Katyusha offers no acceleration over the FISTA on deblurring and Gaussian random inverse problem[1], but significantly outperforms FISTA on the other cases. Indeed, positive results for applying SGD-type algorithms on these problems are well-known already [2, 19, 22], hence we have shown that the SA factor we propose is useful in characterizing whether an inverse problem is inherently a suitable application for stochastic gradient methods.

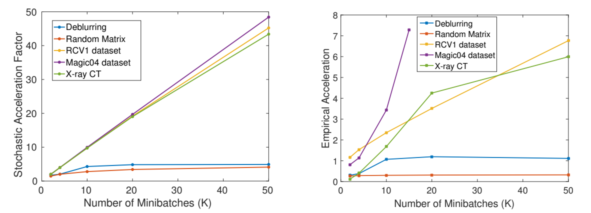

Fig. 2: Left: Stochastic Acceleration (SA) factor of inverse problems with different forward operators, Right: Empirical observation comparing the objective gap convergence of Katyusha and FISTA algorithm in 15 epochs.

4. PRACTICAL ACCELERATION FOR SGD

The previous section suggests that stochastic gradient methods do not always offer an intrinsic advantage for some problems. There are also several other causes for this failure. The most obvious one is that stochastic gradient methods in the primal need to calculate the proximal operator many more times than full gradient methods and hence slow down dramatically the run time. Moreover, in image processing practice often more than one non-smooth regularization term is used, where most of the existing fast stochastic methods such as Katyusha are inapplicable. To avoid the frequent oracle call on the TV proximal operator, we can first reformulate the original optimization problem as a convex-concave saddle-point form. To be specific, the given problem:

$$x^\star \in \min_{x \in \mathbb{R}^d} \left\{f(x) + \lambda g(Dx) + \gamma h(x)\right\}, \tag{11}$$

[1]These two instances cover both scenarios where the observations presented by the rows of A have either a local nature or a global nature.

Algorithm 1 Accelerated Primal-Dual SGD (Acc-PD-SGD)

Initialization: $x^0 = v^0 = v^{-1} \in \mathrm{dom}(g)$, the step size sequences $[\alpha]$, $[\eta]$, $[\theta]$, $l = 0$.
for $t = 1$ **to** N **do**
  $x^t \leftarrow \frac{(3t-2)v^{t-1} + tx^{t-1} - (2t-4)v^{t-2}}{2t+2}$, $x_0 \leftarrow x^t$,
  $z_0 \leftarrow x^t$, $y_0 \quad Dx_0$ → Katyusha-X Momentum
  for $k = 0$ **to** $K - 1$ **do**
    $l \quad l + 1$
    $y_{k+1} = \mathrm{prox}_{\lambda g^*}^{\alpha_l}(y_k + \alpha_l D z_k)$ → Dual Ascent
    Pick $i \in [1, 2, ...K]$ uniformly at random
    $\nabla_k = \nabla f_{S_i}(x_k)$;
    $x_{k+1} = \mathrm{prox}_{\gamma h}^{\eta_l}\left(x_k - \eta_l(D^T y_{k+1} + \nabla_k)\right)$
    → Primal Descent
    $z_{k+1} = x_{k+1} + \theta_l(x_{k+1} - x_k)$
    → Innerloop Momentum
  end for
  $v^t \quad x_K$
end for
Output: x^t

where $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(a_i, x)$ is the data-fidelity term, $g(Dx)$ is a regularization term with a linear operator – for example the TV regularization ($g(.) = \|.\|_1$, $D \in \mathbb{R}^{r\times d}$ is the differential operator), and $h(x)$ is a second convex regularizer. The saddle-point formulation can be written as:

$$[x^\star, y^\star] = \min_{x\in\mathbb{R}^d} \max_{y\in\mathbb{R}^r} f(x) + h(x) + y^T Dx - \lambda g^*(y) \quad (12)$$

The most famous algorithm for solving this saddle-point problem is the Chambolle-Pock algorithm (also known as PDHG) [23, 24], which interleaves the update of the primal variable x and the dual variable y throughout the iterates. With this reformulation the linear operator D and the function $g(.)$ are decoupled and hence one can divide-and-conquer the expensive TV-proximal operator with the primal-dual gradient methods. The stochastic variant of the PDHG for the saddle-point problem (12) has been very recently proposed by Zhao & Cevher [25, Alg.1, "SPDTCM"] and shown to have state-of-the-art performance when compared to PDHG, stochastic ADMM [26] and stochastic proximal averaging [27].

Additionally, since the effect of acceleration given by Nesterov's momentum appears to be very important (for instance, in the experiment from last section, the non-accelerated methods like SVRG perform badly compared to all the accelerated methods), we also need to consider a way to ensure that our method is accelerated. Since the SPDTCM method does not have Nesterov-type acceleration, we propose a variant of it which adopts the outerloop acceleration scheme given by the Katyusha-X algorithm [28]. We observe that such a momentum step is important for the stochastic primal-dual methods in this application. We present our method as Algorithm 1. One can directly choose the same step-size sequences $[\alpha]$, $[\eta]$, $[\theta]$ as suggested in [25, Section 2.3].

We test our algorithm and compare with FISTA [17] and the SPDTCM [25] on a space-varying deblurring task for images sized 512 by 512, with a space-varying out-of-focus blur kernel, and TV-regularization. All algorithm are initialized with a backprojection. We use a machine with 1.6 GB RAM, 2.60 GHz Intel Core i7-5600U CPU and MATLAB R2015b.

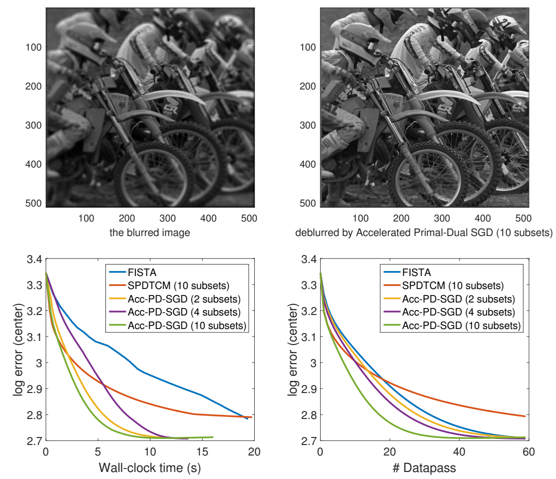

Fig. 3: The estimation error plot for the deblurring experiment with TV-regularization. Image: Kodim05 [16], with an additive Guassian noise (variance 1).

We observe improvement in run time comparing to FISTA since our algorithm can avoid the heavy cost of the TV proximal operator while maintaining the fast convergence provided by Nesterov-type momentum and randomization. We also report a significant convergence improvement over the SPDTCM algorithm both in time and iteration complexity.

5. CONCLUSION

In this work we investigated the value of the state-of-the-art stochastic gradient methods in imaging inverse problems where we chose image deblurring as a running example. We firstly reveal a surprisingly negative result on existing SGD-type methods, and propose a metric (SA) to explain such failures and evaluate the possible computational advantage of using stochastic techniques for a given task; finally we combine several practical ideas and propose the Accelerated Primal-Dual SGD to cope with multiple regularizers (potentially) with a linear operator while maintaining the fast convergence, and demonstrate its effectiveness via deblurring experiments.

6. ACKNOWLEDGEMENTS

We acknowledge the support from H2020-MSCA-ITN 642685 (MacSeNet) , ERC Advanced grant 694888, C-SENSE and a Royal Society Wolfson Research Merit Award. We thank Alessandro Foi, Vladimir Katkovnik, Cristovao Cruz, Enrique Sanchez-Monge, Zhongwei Xu, Alessandro Perelli, Jonathan Mason and Mohammad Golbabaee for helpful discussions. We also thank anonymous reviewers for helpful comments.

7. REFERENCES

[1] Herbert Robbins and Sutton Monro, "A stochastic approximation method," in *Herbert Robbins Selected Papers*, pp. 102–109. Springer, 1985.

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

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

[4] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Advances in Neural Information Processing Systems 26*, pp. 315–323. Curran Associates, Inc., 2013.

[5] 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.

[6] Yurii Nesterov, "A method of solving a convex programming problem with convergence rate o (1/k2)," in *Soviet Mathematics Doklady*, 1983, vol. 27, pp. 372–376.

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

[8] Guanghui Lan and Yi Zhou, "An optimal randomized incremental gradient method," *arXiv preprint arXiv:1507.02000*, 2015.

[9] Zeyuan Allen-Zhu, "Katyusha: The first direct acceleration of stochastic gradient methods," *Journal of Machine Learning Research*, vol. 18, no. 221, pp. 1–51, 2018.

[10] Tomoya Murata and Taiji Suzuki, "Doubly accelerated stochastic variance reduced dual averaging method for regularized empirical risk minimization," in *Advances in Neural Information Processing Systems*, 2017, pp. 608–617.

[11] Junqi Tang, Mohammad Golbabaee, Francis Bach, and Mike E Davies, "Rest-katyusha: Exploiting the solutions structure via scheduled restart schemes," in *Advances in Neural Information Processing Systems 31*, pp. 427–438. Curran Associates, Inc., 2018.

[12] Yu Sun, Brendt Wohlberg, and Ulugbek S Kamilov, "An online plug-and-play algorithm for regularized image reconstruction," *IEEE Transactions on Computational Imaging*, 2019.

[13] Davood Karimi and Rabab 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, pp. 015008, 2016.

[14] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schonlieb, "Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications," *SIAM Journal on Optimization*, vol. 28, no. 4, pp. 2783–2808, 2018.

[15] Oliver Whyte, Josef Sivic, Andrew Zisserman, and Jean Ponce, "Non-uniform deblurring for shaken images," *International journal of computer vision*, vol. 98, no. 2, pp. 168–186, 2012.

[16] Kodak-Lossless-True-Color-Image-Suite, "http://r0k.us/graphics/kodak/," .

[17] A. Beck and M. Teboulle, "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.

[18] Guanghui Lan, "An optimal method for stochastic composite optimization," *Mathematical Programming*, vol. 133, no. 1-2, pp. 365–397, 2012.

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

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

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

[22] Hakan Erdogan and Jeffrey A Fessler, "Ordered subsets algorithms for transmission tomography," *Physics in Medicine & Biology*, vol. 44, no. 11, pp. 2835, 1999.

[23] Antonin Chambolle and Thomas 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.

[24] Antonin Chambolle and Thomas Pock, "On the ergodic convergence rates of a first-order primal–dual algorithm," *Mathematical Programming*, vol. 159, no. 1-2, pp. 253–287, 2016.

[25] Renbo Zhao and Volkan Cevher, "Stochastic three-composite convex minimization with a linear operator," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 765–774.

[26] Hua Ouyang, Niao He, Long Tran, and Alexander Gray, "Stochastic alternating direction method of multipliers," in *International Conference on Machine Learning*, 2013, pp. 80–88.

[27] Wenliang Zhong and James Kwok, "Accelerated stochastic gradient method for composite regularization," in *Artificial Intelligence and Statistics*, 2014, pp. 1086–1094.

[28] Zeyuan Allen-Zhu, "Katyusha x: Practical momentum method for stochastic sum-of-nonconvex optimization," *arXiv preprint arXiv:1802.03866*, 2018.