Stochastic Primal-Dual Three Operator Splitting with Arbitrary Sampling and Preconditioning

Junqi Tang, Matthias Ehrhardt, Carola-Bibiane Schönlieb JT814@CAM.AC.UK
DAMTP, University of Cambridge

Abstract

In this work we propose a stochastic primal-dual preconditioned three-operator splitting algorithm for solving a class of convex three-composite optimization problems. Our proposed scheme is a direct three-operator splitting extension of the SPDHG algorithm [Chambolle et al. 2018]. We provide theoretical convergence analysis showing ergodic $O(1/K)$ convergence rate, and demonstrate the effectiveness of our approach in imaging inverse problems.

1. Introduction

In many imaging inverse problems and machine learning applications, we often need to solve convex composite optimization problems of the form:

$$x^\star = \min_{x \in X} f(Ax) + g(x) + h(x) \tag{1}$$

where $X \in \mathbb{R}^d$ is a convex set, $f(Ax)$ the data-fidelity, $g(x)$ a regularizer which admits simple proximal operator, and $h(x)$ another regularizer for which we have access of its gradients. Here we assume that both three terms are proper convex lower-semicontinuous functions, while $h(x)$ has L-Lipschitz continuous gradients. Since this type of optimization problems are usually large-scale and high-dimensional for modern machine learning and imaging tasks, fast and accurate iterative algorithms are highly desirable in real world applications (Xiao and Zhang, 2014; Defazio et al., 2014; Allen-Zhu, 2017; Tang et al., 2017). The generic composite optimization problem (1) requires three operator splitting schemes as the optimizer (Raguet et al., 2013; Davis and Yin, 2017). Since for most applications the linear operator A is non-trivial, for the case where f is non-smooth (or have a large gradient Lipschitz constant), it is often more beneficial to dualize $f(Ax)$ and reformulate the objective as the saddle-point problem:

$$[x^\star, y^\star] = \min_{x \in X} \max_{y \in Y} \Phi(x, y) := h(x) + g(x) + \langle Ax, y \rangle - f^*(y), \tag{2}$$

where:

$$f^*(y) = \sup_z \langle z, y \rangle - f(z). \tag{3}$$

A number of primal-dual first-order algorithms have been proposed for solving the saddle point problem (2), and among this line of works, the most noticeable one would be the Condat-Vu's scheme (Condat, 2013; Vũ, 2013). In this work, based on the framework of the stochastic primal-dual hybrid gradient (SPDHG) by Chambolle et al. (2018), we propose a stochastic gradient extension of the Condat-Vu algorithm for efficiently solving the saddle point optimization problem (2) if it admits finite-sum structure which is true for most applications in machine learning and computational imaging.

2. Stochastic primal-dual hybrid gradient and three operator splitting

We start by recalling the stochastic primal-dual hybrid gradient (SPDHG) algorithm proposed by Chambolle et al. (2018) for two-composite problems. For simplicity we illustrate here their uniform serial sampling $p_i = p = 1/n$ case:

$$f(y) = \sum_{i=1}^{n} f_i(y_i), \;\; (Ax)_i = A_i x. \tag{4}$$

Now we can formulate the objective to the finite-sum saddle-point form:

$$[x^\star, y^\star] = \min_{x \in X} \max_{y \in Y} \Phi(x, y) := h(x) + g(x) + \sum_{i=1}^{n} \langle A_i x, y_i \rangle - f_i^*(y_i). \tag{5}$$

The most basic version of SPDHG algorithm (Chambolle et al., 2018) for the case $h(x) = 0$ can be written as:

$$x^{k+1} = \mathrm{prox}_g^{\tau}(x^k - \tau z^k) \tag{6}$$

$$y_i^{k+1} = \begin{cases} \mathrm{prox}_{f_i^*}^{\sigma}(y_i^k + \sigma A_i x^{k+1}) & \text{if } i = j_k \\ y_i^k \end{cases} \tag{7}$$

$$z^{k+1} = A^* y^{k+1} + p^{-1} A_{j_k}^* (y_i^{k+1} - y_i^k) = A^* y^k + (p^{-1} + 1) A_{j_k}^* (y_i^{k+1} - y_i^k) \tag{8}$$

where the proximal operator is defined as:

$$\mathrm{prox}_g^{\tau}(x) := \arg\min_{z \in C} \frac{1}{2} \|z - x\|_{\tau^{-1}}^2 + g(z) \tag{9}$$

where the weighted norm $\|x\|_{\tau^{-1}} = \langle \tau^{-1} x, x \rangle$.

Now for the case where we additionally have a smooth regularization term $h(x)$ for which we can compute the gradient but not necessarily the proximal operator, we propose TOS-SPDHG (Three operator splitting with stochastic primal-dual hybrid gradient), which is a simple extension of the SPDHG algorithm:

TOS-SPDHG – Initialize $x^0 \in \mathbb{R}^d$ $y^0 \in \mathbb{R}^q$

For $\;\; k = 0, 1, 2, ..., K$

$$\left| \begin{array}{l} \text{Select } j_k \in \{1, ..., n\} \\ x^{k+1} = \mathrm{prox}_g^{T_k}(x^k - T_k(A^T \bar{y}^k + \nabla h(x^k))) \\ y_i^{k+1} = \begin{cases} \mathrm{prox}_{f_i^*}^{S_k^i}(y_i^k + S_k^i A_i x^{k+1}) & \text{if } i = j_k \\ y_i^k \end{cases} \\ \bar{y}^{k+1} = y^{k+1} + \theta Q(y^{k+1} - y^k) \end{array} \right.$$

Here we denote $S_k = \mathrm{diag}(S_k^1, ..., S_k^n)$ and T_k the step size matrices[1], and scalar $\theta > 0$ while $Q := \mathrm{diag}(p_1^{-1} I, ..., p_n^{-1} I)$ where each of the p_i denotes the probability i-th index is sampled in each iteration.

1. our current analysis in this note restricts only to the case of fixed step size matrices $S_k = S, T_k = T$ but in our ongoing works we found in practice that these can be varying and adaptively chosen for faster rates.

Although this extension of SPDHG is simple and natural, we believe that it is a crucial extension especially for imaging inverse problems. For example, for Positron Emission Tomography imaging tasks, practitioners typically use jointly the non-smooth Kullback-Leibler (KL) data-fit and the relative difference prior (the Q.Clear framework by GE Healthcare) which is a smooth regularizer with simple gradient evaluation but inefficient proximal evaluation, to cope with the extremely noisy measurement data. In such case the original SPDHG is not directly applicable and we need this TOS-SPDHG extension.

Noting that just as the original SPDHG, we keep the non-scalar step-sizes T, S for TOS-SPDHG which can be chosen as preconditioners for accelerating empirical convergence (Ehrhardt et al., 2017). Our convergence theory is built upon the SPDHG analysis by Chambolle et al. (2018) and hence we also support arbitrary sampling schemes, for example we might use importance sampling to accelerate convergence in some applications.

3. Convergence Theory

We start by introducing the definition of expected separable overapproximation (ESO) property (Richtárik and Takáč, 2014) which is standard for the analysis of coordinate descent (Fercoq and Richtárik, 2015), stochastic variance-reduced gradient methods (Konečnỳ et al., 2016), and stochastic primal-dual gradient methods (Chambolle et al., 2018) with arbitrary sampling:

Definition 1 *(Richtárik and Takáč, 2014) (Expected Separable Overapproximation) Let* $\mathbb{S} \in \{1, ..., n\}$ *and* $p_i := \mathbb{P}(i \in \mathbb{S})$*, and bounded linear operators* C_i *and* $(Cx)_i := C_i x$*, then the ESO parameters* $\{v_i\}$ *satisfy:*

$$\mathbb{E}_{\mathbb{S}} \| \sum_{i \in \mathbb{S}} C_i^T z_i \|_2^2 \leq \sum_{i=1}^{n} p_i v_i \|z_i\|_2^2 \tag{10}$$

As we will see, the choice of T, S for provable convergence of TOS-SPDHG depends on the ESO parameters of the linear operator $S^{\frac{1}{2}} A (T^{-1} - LI)^{-\frac{1}{2}}$, where L is the gradient Lipschitz constant of h.

3.1 Convergence analysis

For the convergence proof of TOS-SPDHG algorithm we need following two lemmas.

Lemma 2 *Assuming that h has L-Lipschitz continuous gradients, and both f, g, and h are proper convex lower-semicontinous functions, for the deterministic updates*

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_g^{T_k}(x^k - T_k(A^T \bar{y}^k + \nabla h(x^k))) \\ \hat{y}_i^{k+1} &= \operatorname{prox}_{f_i^*}^{S_k^i}(y_i^k + S_k^i A_i x^{k+1}), \quad i = 1, ..., n, \end{aligned} \tag{11}$$

we have:

$$\begin{aligned} &\|x^k - x\|_{T_k^{-1}}^2 + \|y^k - y\|_{S_k^{-1}}^2 \\ \geq &\|x^{k+1} - x\|_{T_k^{-1}} + \|\hat{y}^{k+1} - y\|_{S_k^{-1}} \\ &+ 2\left(g(x^{k+1}) - g(x) + h(x^{k+1}) - h(x) + f^*(\hat{y}^{k+1}) - f^*(y)\right) - 2\langle A(x - x^{k+1}), \bar{y}^k \rangle \\ &+ \|x^{k+1} - x^k\|_{T_k^{-1} - LI}^2 + \|y^{k+1} - y^k\|_{S_k^{-1}}^2 \end{aligned} \tag{12}$$

We include the proof in the first half of next subsection. The following Lemma can be directly derived from (Chambolle et al., 2018, Lemma 4.2):

Lemma 3 *Let $\mathbb{S} \in 1, ..., n$ be a random set, $y_i^+ = \hat{y}_i$ if $i \in \mathbb{S}$ and y_i otherwise, and $\{v_i\}$ be some ESO parameter of $S^{\frac{1}{2}} A (T^{-1} - LI)^{-\frac{1}{2}}$, then for any $c > 0$ we have:*

$$2\mathbb{E}_{\mathbb{S}} \langle QAx, y^+ - y \rangle \geq -\mathbb{E}_{\mathbb{S}} \left\{ \frac{1}{c} \|x\|_{T^{-1}-LI} + c \max_i \frac{v_i}{p_i} \|y^+ - y\|_{QS^{-1}}^2 \right\} \tag{13}$$

Now following similar steps of the proof for (Chambolle et al., 2018, Theorem 4.3), with a different step-size choice taking into account the L-smoothness of h, we can have the following convergence guarantee for TOS-SPDHG:

Theorem 4 *Assuming that h has L-Lipschitz continuous gradients, and both f, g, and h are proper convex lower-semicontinous functions, let $\theta = 1$ and fixed step-size matrices T, S be selected such that $(T^{-1} - LI)$ is positive-definite and there exist ESO parameters $\{v_i\}$ of $S^{\frac{1}{2}} A (T^{-1} - LI)^{-\frac{1}{2}}$ with $v_i < p_i, \forall i \in [n]$, then we have:*

$$\mathbb{E} \left(\Phi(x_{(K)}, y^\star) - \Phi(x^\star, y_{(K)}) \right) \leq O \left(\frac{1}{K} \right) \tag{14}$$

where $x_{(K)} = \frac{1}{K} \sum_{k=1}^K x^K$, $y_{(K)} = \frac{1}{K} \sum_{k=1}^K y^K$ denotes the ergodic sequences.

3.2 Convergence proof

The proof of the main theorem follows similar steps in (Chambolle et al., 2018) which was originally for two-composite problems, while here we extend their framework to three-composite setting. The first part of the proof is regarding Lemma 2. By the definition of proximal operator we have:

$$x^{k+1} - x^k + T_k(A^T \bar{y} + \nabla h(x^k)) \in T_k g'(x^{k+1}) \tag{15}$$

and due to the convexity of g, for any x we have:

$$g(x) - g(x^{k+1}) - \langle g'(x^{k+1}), x - x^{k+1} \rangle \geq 0, \tag{16}$$

and hence:

$$g(x) \geq g(x^{k+1}) + \langle T_k^{-1}(x^k - x^{k+1}) - A^T \bar{y} - \nabla h(x^k), x - x^{k+1} \rangle \tag{17}$$

Identically, by the definition of proximal operator and convexity of f_i^*, for any y we can have:

$$f_i^*(x) \geq f_i^*(\hat{y}_i^{k+1}) + \langle (S_k^i)^{-1}(y_i^k - \hat{y}_i^{k+1}) - A_i x^{k+1}, y_i - \hat{y}^{k+1} \rangle \tag{18}$$

Summing up the inequalities we have:

$$\begin{aligned} f^*(y) + g(x) \geq\ & f^*(\hat{y}^{k+1}) + g(x^{k+1}) + \|x^k - x^{k+1}\|_{T_k^{-1}}^2 + \|x^{k+1} - x\|_{T_k^{-1}}^2 \\ & - \|x^k - x\|_{T_k^{-1}}^2 - \langle A(x - x^{k+1}), \bar{y}^k \rangle - \langle \nabla h(x^k), x - x^{k+1} \rangle \\ & + \|y^k - \hat{y}^{k+1}\|_{S_k^{-1}}^2 + \|y - \hat{y}^{k+1}\|_{S_k^{-1}}^2 - \|y^k - y\|_{S_k^{-1}}^2 \\ & + \langle Ax^{k+1}, y - \hat{y}^{k+1} \rangle. \end{aligned} \tag{19}$$

Then we arrange the terms to be:

$$
\begin{aligned}
\|x^k - x\|^2_{T_k^{-1}} + \|y^k - y\|^2_{S_k^{-1}} \geq \; & \|x^{k+1} - x\|^2_{T_k^{-1}} + \|y - \hat{y}^{k+1}\|^2_{S_k^{-1}} \\
& + 2(g(x^{k+1}) - g(x) + f^*(\hat{y}^{k+1}) - f^\star(y)) \\
& + 2(\langle Ax^{k+1}, y - \hat{y}^{k+1}\rangle - \langle A(x - x^{k+1}), \bar{y}\rangle) \\
& + \|y^k - \hat{y}^{k+1}\|^2_{S_k^{-1}} + \|x^k - x^{k+1}\|^2_{T_k^{-1}} \\
& - \langle \nabla h(x^k), x - x^{k+1}\rangle.
\end{aligned} \tag{20}
$$

Noting that we have made the assumption that h is L-smooth, hence for this extra term $-\langle \nabla h(x^k), x - x^{k+1}\rangle$ we can bound it using the three-point descent identity for convex and smooth functions:

$$
\langle \nabla h(x^k), x^{k+1} - x\rangle \geq h(x^{k+1}) - h(x) - \frac{L}{2}\|x^{k+1} - x^k\|_2^2, \tag{21}
$$

then by summing these two inequalities we finish the proof of Lemma 2:

$$
\begin{aligned}
& \|x^k - x\|^2_{T_k^{-1}} + \|y^k - y\|^2_{S_k^{-1}} \\
\geq & \|x^{k+1} - x\|_{T_k^{-1}} + \|\hat{y}^{k+1} - y\|_{S_k^{-1}} \\
& + 2\left(g(x^{k+1}) - g(x) + h(x^{k+1}) - h(x) + f^*(\hat{y}^{k+1}) - f^*(y)\right) - 2\langle A(x - x^{k+1}), \bar{y}^k\rangle \\
& + \|x^{k+1} - x^k\|^2_{T_k^{-1} - LI} + \|y^{k+1} - y^k\|^2_{S_k^{-1}}
\end{aligned} \tag{22}
$$

Now let's denote $w = (x, y)$, $w' = (x', y')$ and keep the step-size matrices T, S fixed $\forall k$, denote:

$$
F_i(y_i|x', y_i') := f_i^*(y_i) - f_i^*(y_i') - \langle A_i x', y_i - y_i'\rangle, \quad F(y|w') := \sum_{i=1}^n F_i(y_i|x', y_i'), \tag{23}
$$

and:

$$
F^p(y|w') := \sum_{i=1}^n (1/p_i - 1) F_i(y_i|x', y_i'), \tag{24}
$$

and:

$$
G(x|w') := g(x) + h(x) - g(x') - h(x') - \langle -A^T y', x - x'\rangle, \quad H(w|w') := G(x|w') + F(y|w'). \tag{25}
$$

Then we can have:

$$
\begin{aligned}
& \|x^k - x\|^2_{T^{-1}} + \|y^k - y\|^2_{S^{-1}} + 2F^p(y^k|w) \\
\geq & \mathbb{E}\{\|x^{k+1} - x\|^2_{T^{-1}} + \|\hat{y}^{k+1} - y\|^2_{S^{-1}} + 2F^p(y^{k+1}|w) + 2H(w^{k+1}|w) \\
& - 2\langle A(x^{k+1} - x), Q(y^{k+1} - y^k) + y^k - \bar{y}^k\rangle + \|x^{k+1} - x^k\|^2_{T^{-1} - LI} + \|y^{k+1} - y^k\|^2_{QS^{-1}}\}
\end{aligned} \tag{26}
$$

Meanwhile denote $\beta := \max_{i \in [n]} \frac{v_i}{p_i}$, by ESO we can have:

$$
2\mathbb{E}\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle \geq -\mathbb{E}\left\{\beta\|x^{k+1} - x^k\|_{T^{-1} - LI} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\right\}. \tag{27}
$$

Then:

$$
\begin{aligned}
&\|x^k - x\|^2_{T^{-1}} + \|y^k - y\|^2_{QS^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}} + 2F^p(y^k|w) - 2\langle QA(x^k - x), y^k - y^{k-1}\rangle \\
\geq&\mathbb{E}\{\|x^{k+1} - x\|^2_{T^{-1}} + \|y^{k+1} - y\|^2_{S^{-1}} + 2F^p(y^{k+1}|w) + 2H(w^{k+1}|w) \\
&- 2\langle QA(x^{k+1} - x), y^{k+1} - y^k\rangle + 2\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle + \|x^{k+1} - x^k\|^2_{T^{-1}-LI} + \|y^{k+1} - y^k\|^2_{QS^{-1}}\} \\
\geq&\mathbb{E}\{\|x^{k+1} - x\|^2_{T^{-1}} + \|y^{k+1} - y\|^2_{QS^{-1}} + 2F^p(y^{k+1}|w) + 2H(w^{k+1}|w) \\
&- 2\langle QA(x^{k+1} - x), y^{k+1} - y^k\rangle - \beta\|x^{k+1} - x^k\|_{T^{-1}-LI} \\
&+ \|x^{k+1} - x^k\|^2_{T^{-1}-LI} + \|y^{k+1} - y^k\|^2_{QS^{-1}}\}
\end{aligned}
\tag{28}
$$

Let's denote:

$$
\Omega^k := \mathbb{E}\{\|x^k - x\|^2_{T^{-1}} + \|y^k - y\|^2_{QS^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}} + 2F^p(y^k|w) - 2\langle QA(x^k - x), y^k - y^{k-1}\rangle\}, \tag{29}
$$

for $\gamma < 1$ and $T^{-1} - LI$ being positive definite we have:

$$
\Omega^k \geq \Omega^{k+1} + \mathbb{E}(2H(w^{k+1}|w) + (1-\beta)\|x^{k+1} - x^k\|_{T^{-1}+LI}) \geq \Omega^{k+1} + 2\mathbb{E}H(w^{k+1}|w). \tag{30}
$$

By summing over above inequality from $k = 0$ to $K - 1$, at the same time notice that $\Omega^k \geq 2\mathbb{E}F^p(y^k|w)$ due to ESO, we have:

$$
\mathbb{E}\sum_{k=1}^{K} H(w^k|w) \leq \frac{\Omega^0}{2} - \mathbb{E}F^p(y^K|w) < +\infty, \tag{31}
$$

then due to convexity of H, for the ergodic sequence $w_(k) = \frac{1}{K}\sum_{k=1}^{K} w_k$, we have:

$$
\mathbb{E}\left(\Phi(x_{(K)}, y^\star) - \Phi(x^\star, y_{(K)})\right) \leq \mathbb{E}H(w_{(K)}|w) \leq \frac{1}{K}\mathbb{E}\sum_{k=1}^{K} H(w^k|w) = O(1/K) \tag{32}
$$

Thus finishes the proof of the Theorem 4.

4. Numerical Experiment

In this section we present our preliminary numerical experiments on X-ray Computed Tomography (CT) image reconstruction which is an imaging inverse problem suitable for stochastic first-order methods (Tang et al., 2020), where we seek to find an estimate of some ground-truth image $x^\dagger$ via measurement b corrupted by Poisson noise:

$$
b \sim \text{Poisson}(I_0 e^{-Ax^\dagger}), \tag{33}
$$

via solving:

$$
x^\star = \min_{x \in X} f(Ax, b) + g(x) + h(x). \tag{34}
$$

Here we choose the smooth edge-preserving regularizer (Thibault et al., 2007) $h(x) = \lambda \sum_r \phi([Dx]_r)$ where D is the 2D differential operator while the potential function $\phi(z) = \frac{|z|^p}{1+|\frac{z}{c}|^{p-q}}$ with recommended choice $p = 2, q = 1.5, c = 10$; the second regularizer $g(x) = \iota_C(x)$ to be the indicator function of a box

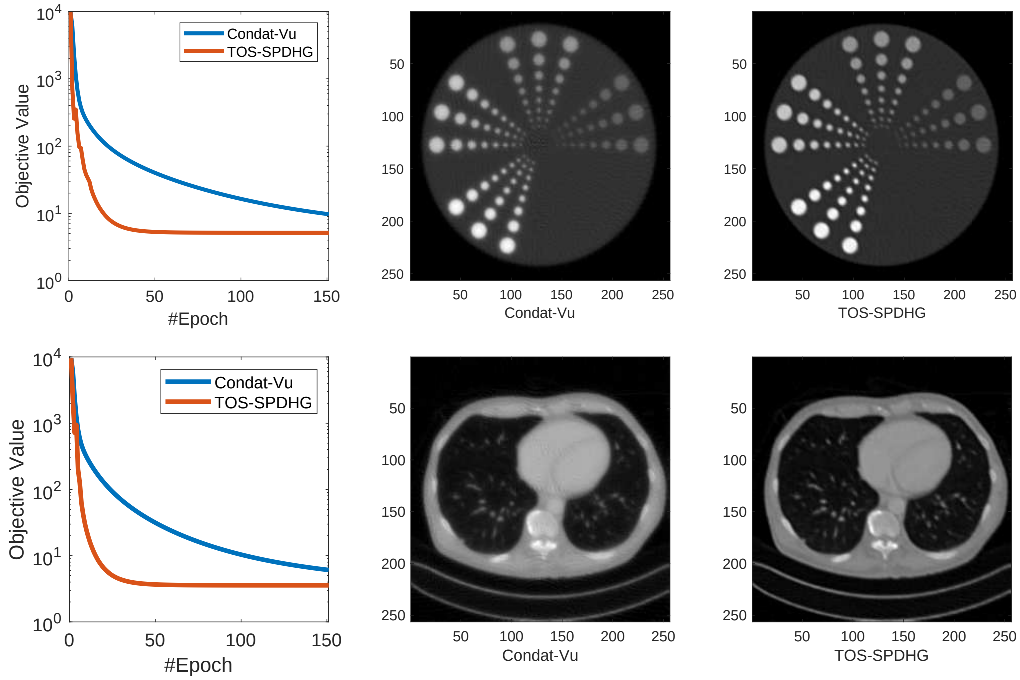

Figure 1: Sparse-view X-ray CT Reconstruction results for Condat-Vu and TOS-SPDHG, termintate at 150th epoch

constraint restricting the pixel values to sit in the range of $[0, 1]$. We consider two CT modelities – sparse-view and low-dose CT, for the first one we choose least-squares datafit which is smooth (for the proof of concept), while for the second one we choose Kullback-Leibler (KL) divergence which is non-smooth. We consider the Condat-Vu's three-operator splitting scheme (Condat, 2013; Vũ, 2013; Chambolle and Pock, 2016) as the baseline – this algorithm can be view as the deterministic counterpart of our algorithm. In both experiments we found that our algorithm converges significantly faster than the Condat-Vu's method in this type of three-composite problems with at least $O(1/K)$ rates, complimenting our theory.

5. Conclusion

In this note we propose a stochastic primal-dual three operator splitting algorithm TOS-SPDHG admitting preconditioning and arbitrary sampling, for efficiently solving convex three-composite optimization problems, particularly in imaging applications. We build upon the theoretical framework of (Chambolle et al., 2018) and extend the convergence proof to three-composite case, showing ergodic $O(1/K)$ convergence rate. Moreover, if any part of the function, either f^*, g or h is strongly-convex, we can trivially extend the current theory to prove faster convergence rates, which we omit here for now. Our algorithm is a simple yet crucial extension of SPDHG, with a much wider range of applications, especially in computational and medical imaging tasks where non-smooth data-fidelity and smooth regularizers are recommended, such as clinical 3D PET reconstruction.

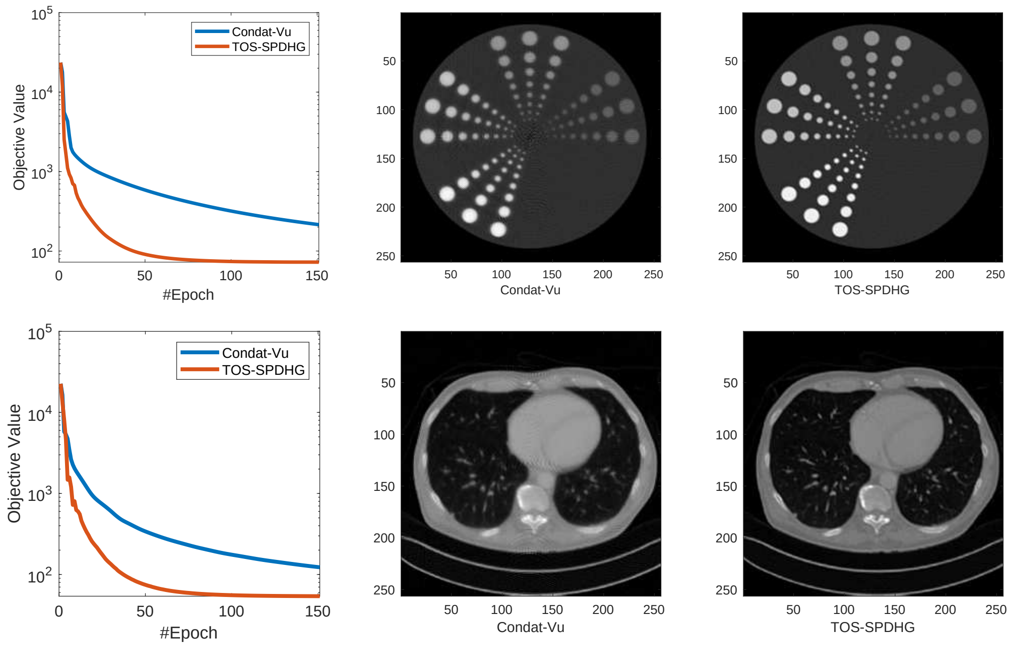

Figure 2: Low-dose CT Reconstruction results for Condat-Vu and TOS-SPDHG, termintate at 150th epoch

References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1-2):253–287, 2016.

Antonin Chambolle, Matthias J Ehrhardt, Peter Richtarik, and Carola-Bibiane Schonlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

Laurent Condat. A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. *Journal of optimization theory and applications*, 158(2):460–479, 2013.

Damek Davis and Wotao Yin. A three-operator splitting scheme and its optimization applications. *Set-valued and variational analysis*, 25(4):829–858, 2017.

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*, pages 1646–1654, 2014.

Matthias J Ehrhardt, Pawel Markiewicz, Antonin Chambolle, Peter Richtárik, Jonathan Schott, and Carola-Bibiane Schönlieb. Faster pet reconstruction with a stochastic primal-dual hybrid gradient method. In *Wavelets and Sparsity XVII*, volume 10394, page 103941O. International Society for Optics and Photonics, 2017.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Jakub Konečnỳ, Jie Liu, Peter Richtárik, and Martin Takáč. Mini-batch semi-stochastic gradient descent in the proximal setting. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):242–255, 2016.

Hugo Raguet, Jalal Fadili, and Gabriel Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

Junqi Tang, Mohammad Golbabaee, and Mike E. Davies. Gradient projection iterative sketch for large-scale constrained least-squares. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3377–3386. PMLR, 2017.

Junqi Tang, Karen Egiazarian, Mohammad Golbabaee, and Mike Davies. The practicality of stochastic optimization in imaging inverse problems. *IEEE Transactions on Computational Imaging*, 6:1471–1485, 2020.

Jean-Baptiste Thibault, Ken D Sauer, Charles A Bouman, and Jiang Hsieh. A three-dimensional statistical approach to improved image quality for multislice helical ct. *Medical physics*, 34(11):4526–4544, 2007.

Bang Công Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.