Noname manuscript No.
(will be inserted by the editor)

A Primal-dual Three-operator Splitting Scheme

Ming Yan

Received: date / Accepted: date

Abstract In this paper, we propose a new primal-dual algorithm for minimizing $f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$, where f, g, and h are convex functions, f is differentiable with a Lipschitz continuous gradient, and $\mathbf{A}$ is a bounded linear operator. It has some famous primal-dual algorithms for minimizing the sum of two functions as special cases. For example, it reduces to the Chambolle-Pock algorithm when $f = 0$ and a primal-dual fixed-point algorithm in [P. Chen, J. Huang, and X. Zhang, A primal-dual fixed-point algorithm for convex separable minimization with applications to image restoration, Inverse Problems, 29 (2013), p.025011] when $g = 0$. In addition, it recovers the three-operator splitting scheme in [D. Davis and W. Yin, A three-operator splitting scheme and its optimization applications, arXiv:1504.01032, (2015)] when $\mathbf{A}$ is the identity operator. We prove the convergence of this new algorithm for the general case by showing that the iteration is a nonexpansive operator and derive the linear convergence rate with additional assumptions. Comparing to other primal-dual algorithms for solving the same problem, this algorithm extends the range of acceptable parameters to ensure the convergence and has a smaller per-iteration cost. The numerical experiments show the efficiency of this new algorithm by comparing to other primal-dual algorithms.

Keywords fixed-point iteration · nonexpansive operator · primal-dual · three-operator splitting

Mathematics Subject Classification (2000) 47H05 · 65K05 · 65K15 · 90C25

The work was supported in part by the NSF grant DMS-1621798.

M. Yan
Department of Computational Mathematics, Science and Engineering
Department of Mathematics
Michigan State University, East Lansing, MI, USA
E-mail: yanm@math.msu.edu

1 Introduction

This paper focuses on minimizing the sum of three proper lower semi-continuous convex functions in the form of

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x}), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are two Hilbert spaces, $f : \mathcal{X} \mapsto \mathbf{R}$ is differentiable with a $1/\beta$-Lipschitz continuous gradient for some $\beta \in (0, +\infty)$, and $\mathbf{A} : \mathcal{X} \mapsto \mathcal{Y}$ is a bounded linear operator. A wide range of problems in image and signal processing, statistic and machine learning can be formulated into this form, and we give some examples.

Elastic net regularization [27]: The elastic net combines the ℓ_1 and ℓ_2 penalties to overcome the limitations of both penalties. The optimization problem is

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \mu_2\|\mathbf{x}\|_2^2 + \mu_1\|\mathbf{x}\|_1 + l(\mathbf{A}\mathbf{x}, \mathbf{b}),$$

where $\mathbf{x} \in \mathbf{R}^p$, $\mathbf{A} \in \mathbf{R}^{n\times p}$, $\mathbf{b} \in \mathbf{R}^n$, and l is the loss function, which may be nondifferentiable. The ℓ_2 regularization term $\mu_2\|\mathbf{x}\|_2^2$ is differentiable and has a Lipschitz continuous gradient.

Fused lasso [23]: The fused lasso was proposed for group variable selection. Except the ℓ_1 penalty, it includes a new penalty term for large changes with respect to the temporal or spatial structure such that the coefficients vary in a smooth fashion. The optimization problem for fused lasso with the least squares loss is

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \mu_1\|\mathbf{x}\|_1 + \mu_2\|\mathbf{D}\mathbf{x}\|_1,$$

where $\mathbf{x} \in \mathbf{R}^p$, $\mathbf{A} \in \mathbf{R}^{n\times p}$, $\mathbf{b} \in \mathbf{R}^n$, and

$$\mathbf{D} = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \cdots & \cdots \\ & & & -1\ 1 \end{pmatrix}$$

is a matrix in $\mathbf{R}^{(p-1)\times p}$.

Image restoration with two regularizations: Many image processing problems have two or more regularizations. For instance, in computed tomorgraph reconstruction, nonnegative constraint and total variation regularization are applied. The optimization problem can be formulated as

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \mu\|\mathbf{D}\mathbf{x}\|_1 + \iota_C(\mathbf{x}),$$

where $\mathbf{x} \in \mathbf{R}^n$ is the image to be reconstructed, $\mathbf{A} \in \mathbf{R}^{m\times n}$ is the forward projection matrix that maps the image to the sinogram data, $\mathbf{b} \in \mathbf{R}^m$ is the measured sinogram data with noise, $\mathbf{D}$ is a discrete gradient operator, and

ι_C is the indicator function that returns zero if $\mathbf{x} \in C$ (here, C is the set of nonnegative vectors in $\mathbf{R}^n$) and $+\infty$ otherwise.

Before introducing the algorithms for solving (1), we discuss special cases of (1) with only two functions. When either f or g is missing, the problem (1) reduces to the sum of two functions, and many splitting and proximal algorithms are proposed and studied in the literature. Two famous groups of methods are Alternating Direction of Multiplier Method (ADMM) and primal-dual algorithms [15]. ADMM applied on a convex optimization problem was shown to be equivalent to Douglas-Rachford Splitting (DRS) applied on the dual problem [13], and Yan and Yin showed recently in [25] that it is also equivalent to DRS applied on the same primal problem. In fact, there are many different ways in reformulating a problem into the ADMM form such that ADMM can be applied, and among these ways, some are equivalent. However, there will be always a subproblem involving $\mathbf{A}$, and it may not be solved analytically depending on the properties of $\mathbf{A}$ and the way ADMM is applied. On the other hand, primal-dual algorithms only need the operator $\mathbf{A}$ and its adjoint operator $\mathbf{A}^\top$[1]. Thus, it has been applied in a lot of applications because the subproblems would be easy to solve if the proximal mappings for both g and h can be computed easily.

The primal-dual algorithms for two and three functions are reviewed in [15] and specially for image processing problems in [5]. When the differentiable function f is missing, the primal-dual algorithm is the Primal-Dual Hybrid Gradient (PDHG or Chambolle-Pock) algorithm [22,12,4], while the primal-dual algorithm (Primal-Dual Fixed-Point algorithm based on the Proximity Operator or PDFP^{2}O) with g missing is proposed in [18,6]. In order to solve the problem (1) with three functions, we can reformulate the problem and apply the primal-dual algorithms for two functions. E.g., we can let $\bar{h}([\mathbf{I};\ \mathbf{A}]\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$ and apply the PDFP^{2}O or let $\bar{g}(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ and apply the Chambolle-Pock algorithm. However, the first approach introduces more dual variables. For the second approach, the proximal mapping of $\bar{g}$ may not be easy to compute and the differentiability of f is not utilized.

When all three terms are considered, Condat and Vu proposed a primal-dual splitting scheme for (1) in [8,24]. It is a generalization of the Chambolle-Pock algorithm by involving the differentiable function f with a more restrictive range for acceptable parameters than the Chambolle-Pock algorithm because of the additional function. Then the Asymmetric Forward-Backward-Adjoint splitting (AFBA) is proposed in [17], and the Condat-Vu algorithm is a special case of this proposed algorithm. As noted in [15], there is no generalization of PDFP^{2}O for three functions at that time. However, a generalization of PDFP^{2}O–a Primal-Dual Fixed-Point algorithm (PDFP)–is proposed in [7], in which two proximal operators of g are needed in every iteration. This algorithm has a larger range of acceptable parameters than the Condat-Vu algorithm. In this paper, we will give a new generalization of both the Chambolle-Pock algorithm and PDFP^{2}O. This new algorithm employs

[1] The adjoint operator $\mathbf{A}^\top$ is defined by $\langle \mathbf{s}, \mathbf{A}\mathbf{x}\rangle_{\mathcal{Y}} = \langle \mathbf{A}^\top \mathbf{s}, \mathbf{x}\rangle_{\mathcal{X}}$

the same regions of acceptable parameters with PDFP and the same per-iteration complexity as the Condat-Vu algorithm. In addition, when $\mathbf{A} = \mathbf{I}$, we recover the three-operator splitting scheme developed by Davis and Yin in [10]. The three-operator splitting in [10] is a generalization of many existing two-operator splitting schemes such as forward-backward splitting [19], backward-forward splitting [21,2], Peaceman-Rachford splitting [11,20], and forward-Douglas-Rachford splitting [3].

The proposed algorithm PD3O has the following iteration ($(\mathbf{z},\mathbf{s}) \to (\mathbf{z}^+,\mathbf{s}^+)$):

$$\mathbf{x} \in (\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z}); \tag{2a}$$

$$\begin{aligned}\mathbf{s}^+ &\in \left(\mathbf{I}+\tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left((\mathbf{I}-\lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s}+\tfrac{\lambda}{\gamma}\mathbf{A}\left(2\mathbf{x}-\mathbf{z}-\gamma\nabla f(\mathbf{x})\right)\right) \\ &= \tfrac{\lambda}{\gamma}(\mathbf{I}-\mathbf{prox}_{\frac{\gamma}{\lambda}h})\left(\tfrac{\gamma}{\lambda}(\mathbf{I}-\lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s}+\mathbf{A}\left(2\mathbf{x}-\mathbf{z}-\gamma\nabla f(\mathbf{x})\right)\right);\end{aligned} \tag{2b}$$

$$\mathbf{z}^+ = \mathbf{x}-\gamma\nabla f(\mathbf{x})-\gamma\mathbf{A}^\top\mathbf{s}^+. \tag{2c}$$

Here the proximal mapping is defined as $\mathbf{prox}_{\gamma g}(\mathbf{z}) \equiv (\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z}) := \arg\min_{\mathbf{x}} \gamma g(\mathbf{x})+\frac{1}{2}\|\mathbf{x}-\mathbf{z}\|^2$. Because the optimization problem for the proximal mapping may have multiple solutions, we have "$\in$" instead of "$=$" in (2a) and (2b). The contributions of this paper can be summarized as follows:

- We proposed a new primal-dual algorithm for solving an optimization problem with three functions $f(\mathbf{x})+g(\mathbf{x})+h(\mathbf{A}\mathbf{x})$ that recovers the Chambolle-Pock algorithm [4] and PDFP^{2}O [6] for two functions with either f or g missing. Though there are two algorithms for solving the same problem: Condat-Vu [8,24] and PDFP [7], this new algorithm combines the advantages of both methods: the low per-iteration complexity of Condat-Vu and the large range of acceptable parameters for convergence of PDFP. The numerical experiments show the advantage of the proposed algorithm over both existing algorithms.
- We prove the convergence of the algorithm by showing that the iteration is an α-averaged operator. This result is stronger than the result for PDFP^{2}O in [6], where the iteration is shown to be nonexpansive only. Also, we show that the Chambolle-Pock algorithm is firmly nonexpansive under a different metric from the previous result that it is equivalent to a proximal point algorithm applied on the Karush-Kuhn-Tucker (KKT) conditions.
- This new algorithm also recovers the recently proposed three-operator splitting by Davis and Yin [10] and thus many splitting schemes involving two operators such as forward-backward splitting, backward-forward splitting, Peaceman-Rachford splitting, and forward-Douglas-Rachford splitting.
- With additional assumptions on the functions, we show the linear convergence rate of this new algorithm.

The rest of the paper is organized as follows. We compare the new algorithm (2) with existing primal-dual algorithms and the three-operator splitting

scheme in Section 2. Then we show the convergence of (2) for the general case and its linear convergence rate in Section 3. The numerical experiments in Section 4 show the effectiveness and efficiency of the proposed algorithm by comparing to other existing algorithms, and finally, Section 5 concludes the paper with future directions.

2 Connections to Existing Algorithms

In this section, we compare our proposed algorithm with several existing algorithms. In particular, we show that our proposed algorithm recovers PDFP^{2}O [18, 6], the Chambolle-Pock algorithm [4], and the three-operator splitting by Davis and Yin [10]. In addition, we compare our algorithm with PDFP [7] and the Condat-Vu algorithm [8,24].

Before showing the connections, we reformulate our algorithm by changing the update order of the variables and introducing $\bar{\mathbf{x}}$ to replace $\mathbf{z}$ (i.e., $\bar{\mathbf{x}} = 2\mathbf{x} - \mathbf{z} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}$). The reformulated algorithm is

$$\mathbf{s}^+ = \left(\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left(\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\bar{\mathbf{x}}\right); \tag{3a}$$

$$\mathbf{x}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}^+); \tag{3b}$$

$$\bar{\mathbf{x}}^+ = 2\mathbf{x}^+ - \mathbf{x} + \gamma\nabla f(\mathbf{x}) - \gamma\nabla f(\mathbf{x}^+). \tag{3c}$$

2.1 PDFP^{2}O

When $g = 0$, i.e., the function g is missing, we have $\mathbf{x} = \mathbf{z}$, and the iteration (2) reduces to

$$\mathbf{s}^+ = \left(\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left((\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\left(\mathbf{x} - \gamma\nabla f(\mathbf{x})\right)\right); \tag{4a}$$

$$\mathbf{x}^+ = \mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}^+, \tag{4b}$$

which is the PDFP^{2}O in [18,6]. The PDFP^{2}O iteration is shown to be nonexpansive in [6], while we will show that PFDP^{2}O, as a special case of our algorithm, is α-averaged with certain $\alpha \in (0, 1)$.

2.2 Chambolle-Pock

Let $f = 0$, i.e., the function f is missing, then we have, from (3),

$$\mathbf{s}^+ = \left(\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left(\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\bar{\mathbf{x}}\right); \tag{5a}$$

$$\mathbf{x}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\mathbf{A}^\top\mathbf{s}^+); \tag{5b}$$

$$\bar{\mathbf{x}}^+ = 2\mathbf{x}^+ - \mathbf{x}, \tag{5c}$$

which is the PDHG in [4]. We will show that it is firmly nonexpansive under a different metric.

2.3 Three-Operator Splitting

Let $\mathbf{A} = \mathbf{I}$ and $\lambda = 1$, then we have, from (2),

$$\mathbf{x} = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{z}); \tag{6a}$$
$$\mathbf{s}^+ = \tfrac{1}{\gamma}(\mathbf{I} - \mathbf{prox}_{\gamma h})\left(2\mathbf{x} - \mathbf{z} - \gamma\nabla f(\mathbf{x})\right); \tag{6b}$$
$$\mathbf{z}^+ = \mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{s}^+, \tag{6c}$$

which is equivalent to the three-operator splitting in [10] because $\mathbf{s}^+$ can be eliminated by combining both (6b) and (6c).

2.4 PDFP

The PDFP algorithm [7] is developed as a generalization of PDFP^{2}O. When $g = \iota_C$, the PDFP is reduced to the Preconditioned Alternating Projection Algorithm (PAPA) algorithm proposed in [16]. The PDFP iteration can be expressed as follows:

$$\mathbf{z} = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}); \tag{7a}$$
$$\mathbf{s}^+ = (\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*)^{-1}\left(\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\mathbf{z}\right); \tag{7b}$$
$$\mathbf{x}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}^+). \tag{7c}$$

Note that two proximal mappings of g are needed in each iteration, while our algorithm only needs one.

2.5 Condat-Vu

The Condat-Vu algorithm [8,24] is a generalization of the Chambolle-Pock algorithm for problem (1). The iteration is

$$\mathbf{s}^+ = \left(\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left(\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\bar{\mathbf{x}}\right); \tag{8a}$$
$$\mathbf{x}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}^+); \tag{8b}$$
$$\bar{\mathbf{x}}^+ = 2\mathbf{x}^+ - \mathbf{x}. \tag{8c}$$

The difference between our algorithm and the Condat-Vu algorithm is in the updating of $\bar{\mathbf{x}}$. Because of the difference, our algorithm will be shown to have more freedom than the Condat-Vu algorithm in choosing acceptable parameters.

The parameters for the three primal-dual algorithms solving (1) and the relation between all the mentioned primal-dual algorithms are given in the following table.

	$f \neq 0,\ g \neq 0$	$f = 0$	$g = 0$
PDFP	$\lambda < 1/\lambda_{\max}(\mathbf{A}\mathbf{A}^\top);\ \gamma < 2\beta$		PDFP^{2}O
Condat-Vu	$\lambda \cdot \lambda_{\max}(\mathbf{A}\mathbf{A}^\top) + \gamma/(2\beta) \leq 1$	Chambolle-Pock	
Our Algorithm	$\lambda < 1/\lambda_{\max}(\mathbf{A}\mathbf{A}^\top);\ \gamma < 2\beta$	Chambolle-Pock	PDFP^{2}O

3 The Proposed Primal-dual Three-operator Fixed-point Algorithm

3.1 Notation and Preliminaries

Let $\mathbf{I}$ be the identity operator defined on a Hilbert space, and, for simplicity, we do not specify the space on which it is defined when it is clear from the context. Let $\mathbf{M} = \mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top$. When λ is small enough such that $\rho(\mathbf{s},\mathbf{t}) := \langle \mathbf{s}, \mathbf{M}\mathbf{t}\rangle$ is a distance defined for $\mathbf{s} \in \mathcal{Y}$ and $\mathbf{t} \in \mathcal{Y}$, we denote $\langle \mathbf{s},\mathbf{t}\rangle_{\mathbf{M}} := \langle \mathbf{s}, \mathbf{M}\mathbf{t}\rangle$ and the induced norm by $\|\mathbf{s}\|_{\mathbf{M}} = \sqrt{\langle \mathbf{s},\mathbf{s}\rangle_{\mathbf{M}}}$. Furthermore, by an abuse of notation, we define a norm for $(\mathbf{z},\mathbf{s}) \in \mathcal{X} \times \mathcal{Y}$ as $\|(\mathbf{z},\mathbf{s})\|_{\mathbf{M}} = \sqrt{\|\mathbf{z}\|^2 + \frac{\gamma^2}{\lambda}\|\mathbf{s}\|^2_{\mathbf{M}}}$. Letting $\tilde{\mathbf{z}} = 2\mathbf{x} - \mathbf{z} - \gamma\nabla f(\mathbf{x})$, we have

$$\mathbf{z}^+ = \mathbf{z} - \mathbf{x} + \tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+. \tag{9}$$

Denote the iteration in (2) as $\mathbf{T}$, i.e., $(\mathbf{z}^+,\mathbf{s}^+) = \mathbf{T}(\mathbf{z},\mathbf{s})$ and the relaxed operator $\mathbf{T}_\theta = \theta\mathbf{T} + (1-\theta)\mathbf{I}$. An operator $\tilde{\mathbf{T}}$ is nonexpansive if $\|\tilde{\mathbf{T}}\mathbf{x} - \tilde{\mathbf{T}}\mathbf{y}\| \leq \|\mathbf{x}-\mathbf{y}\|$ for any $\mathbf{x}$ and $\mathbf{y}$. An operator $\tilde{\mathbf{T}}$ is α-averaged for $\alpha \in (0,1]$ if $\|\tilde{\mathbf{T}}\mathbf{x} - \tilde{\mathbf{T}}\mathbf{y}\|^2 \leq \|\mathbf{x}-\mathbf{y}\|^2 - \frac{1-\alpha}{\alpha}\|(\mathbf{I}-\tilde{\mathbf{T}})\mathbf{x} - (\mathbf{I}-\tilde{\mathbf{T}})\mathbf{y}\|^2$; firmly nonexpansive operators are 1/2-averaged, and nonexpansive operators are 1-averaged.

Because $(\mathbf{I}+\gamma\partial g)^{-1}$ is firmly nonexpansive, we have

$$\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{z}-\mathbf{x}-\mathbf{w}+\mathbf{y}\|^2 \leq \|\mathbf{z}-\mathbf{w}\|^2, \tag{10a}$$

$$\|\mathbf{x}-\mathbf{y}\|^2 \leq \langle \mathbf{x}-\mathbf{y}, \mathbf{z}-\mathbf{w}\rangle, \tag{10b}$$

$$\|\mathbf{z}-\mathbf{x}-\mathbf{w}+\mathbf{y}\|^2 \leq \langle \mathbf{z}-\mathbf{x}-\mathbf{w}+\mathbf{y}, \mathbf{z}-\mathbf{w}\rangle, \tag{10c}$$

where $\mathbf{x} = (\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z})$ and $\mathbf{y} = (\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{w})$. In addition, we have that ∇f is β-cocoercive, i.e., $\langle \mathbf{x}-\mathbf{y}, \nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\rangle \geq \beta\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$ for any $\mathbf{x},\ \mathbf{y} \in \mathcal{X}$ because f has a $1/\beta$ Lipschitz continuous gradient [1, Theorem 18.15].

3.2 Convergence Analysis for the General Case

In this section, we show the convergence of the proposed algorithm in Theorem 1. We show firstly that the iteration $\mathbf{T}$ is a nonexpansive operator (Lemma 1) and then finding a fixed point $(\mathbf{z}^*,\mathbf{s}^*)$ of $\mathbf{T}$ is equivalent to finding an optimal solution to (1) (Lemma 2).

Lemma 1 *The iteration $\mathbf{T}$ in (2) is a nonexpansive operator for $(\mathbf{z},\mathbf{s})$ if $\gamma \leq 2\beta$. Furthermore, it is α-averaged with $\alpha = \frac{2\beta}{4\beta-\gamma}$.*

Proof Let $(\mathbf{w}^+, \mathbf{t}^+) = \mathbf{T}(\mathbf{w}, \mathbf{t})$ and

$$\begin{aligned}
\mathbf{y} &\in (\mathbf{I} + \gamma \partial g)^{-1}(\mathbf{w}),\\
\tilde{\mathbf{w}} &= 2\mathbf{y} - \mathbf{w} - \gamma \nabla f(\mathbf{y}),\\
\mathbf{t}^+ &\in \left(\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*\right)^{-1}\left((\mathbf{I} - \lambda \mathbf{A}\mathbf{A}^\top)\mathbf{t} + \tfrac{\lambda}{\gamma}\mathbf{A}\tilde{\mathbf{w}}\right),\\
\mathbf{w}^+ &= \mathbf{y} - \gamma \nabla f(\mathbf{y}) - \gamma \mathbf{A}^\top \mathbf{t}^+.
\end{aligned}$$

Then we have

$$\begin{aligned}
\|\mathbf{z}^+ - \mathbf{w}^+\|^2 =& \|\mathbf{z} - \mathbf{x} + \tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \mathbf{w} + \mathbf{y} - \tilde{\mathbf{w}} + \gamma\mathbf{A}^\top\mathbf{t}^+\|^2\\
=& \|\mathbf{z} - \mathbf{x} - \mathbf{w} + \mathbf{y}\|^2 + \|\tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \tilde{\mathbf{w}} + \gamma\mathbf{A}^\top\mathbf{t}^+\|^2\\
&+ 2\langle \mathbf{z} - \mathbf{x} - \mathbf{w} + \mathbf{y}, \tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \tilde{\mathbf{w}} + \gamma\mathbf{A}^\top\mathbf{t}^+\rangle\\
\le& \langle \mathbf{z} - \mathbf{x} - \mathbf{w} + \mathbf{y}, \mathbf{z} - \mathbf{w}\rangle + \|\tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \tilde{\mathbf{w}} + \gamma\mathbf{A}^\top\mathbf{t}^+\|^2\\
&+ 2\langle \mathbf{z} - \mathbf{x} - \mathbf{w} + \mathbf{y}, \mathbf{z}^+ - \mathbf{z} + \mathbf{x} - \mathbf{w}^+ + \mathbf{w} - \mathbf{y}\rangle.
\end{aligned}$$

Here we used (9) for $\mathbf{z}^+$ and $\mathbf{w}^+$, and the inequality comes from (10c). In addition, we have

$$\begin{aligned}
&\|\tilde{\mathbf{z}} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \tilde{\mathbf{w}} + \gamma\mathbf{A}^\top\mathbf{t}^+\|^2 + \tfrac{\gamma^2}{\lambda}\|\mathbf{s}^+ - \mathbf{t}^+\|_{\mathbf{M}}^2\\
=& \|\tilde{\mathbf{z}} - \tilde{\mathbf{w}}\|^2 - 2\gamma\langle \tilde{\mathbf{z}} - \tilde{\mathbf{w}}, \mathbf{A}^\top\mathbf{s}^+ - \mathbf{A}^\top\mathbf{t}^+\rangle + \tfrac{\gamma^2}{\lambda}\|\mathbf{s}^+ - \mathbf{t}^+\|^2\\
\le& \langle \tilde{\mathbf{z}} - \tilde{\mathbf{w}}, (\tilde{\mathbf{z}} - \tilde{\mathbf{w}}) - 2\gamma(\mathbf{A}^\top\mathbf{s}^+ - \mathbf{A}^\top\mathbf{t}^+)\rangle\\
&+ \tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{t}^+, (\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\tilde{\mathbf{z}} - (\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{t} + \tfrac{\lambda}{\gamma}\mathbf{A}\tilde{\mathbf{w}}\rangle\\
=& \langle \tilde{\mathbf{z}} - \tilde{\mathbf{w}}, (\tilde{\mathbf{z}} - \tilde{\mathbf{w}}) - \gamma(\mathbf{A}^\top\mathbf{s}^+ - \mathbf{A}^\top\mathbf{t}^+)\rangle + \tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{t}^+, (\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)(\mathbf{s} - \mathbf{t})\rangle\\
=& \langle \tilde{\mathbf{z}} - \tilde{\mathbf{w}}, \mathbf{z}^+ - \mathbf{z} + \mathbf{x} - \mathbf{w}^+ + \mathbf{w} - \mathbf{y}\rangle + \tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{t}^+, (\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)(\mathbf{s} - \mathbf{t})\rangle.
\end{aligned}$$

Here the inequality comes from the nonexpansiveness of $(\mathbf{I}+\frac{\lambda}{\gamma}\partial h^*)^{-1}$; see (10b). Combining the previouse two equations, we have

$$\begin{aligned}
&\|(\mathbf{z}^+,\mathbf{s}^+)-(\mathbf{w}^+,\mathbf{t}^+)\|_{\mathbf{M}}^2\\
\leq&\langle \mathbf{z}-\mathbf{x}-\mathbf{w}+\mathbf{y},\mathbf{z}-\mathbf{w}\rangle+\langle\tilde{\mathbf{z}}-\tilde{\mathbf{w}},\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
&+\tfrac{\gamma^2}{\lambda}\langle\mathbf{s}^+-\mathbf{t}^+,(\mathbf{I}-\lambda\mathbf{A}\mathbf{A}^\top)(\mathbf{s}-\mathbf{t})\rangle\\
&+2\langle\mathbf{z}-\mathbf{x}-\mathbf{w}+\mathbf{y},\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
=&\langle\mathbf{z}^+-\mathbf{w}^+,\mathbf{z}-\mathbf{w}\rangle+\langle\tilde{\mathbf{z}}-\tilde{\mathbf{w}},\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
&+\tfrac{\gamma^2}{\lambda}\langle\mathbf{s}^+-\mathbf{t}^+,(\mathbf{I}-\lambda\mathbf{A}\mathbf{A}^\top)(\mathbf{s}-\mathbf{t})\rangle\\
&+\langle\mathbf{z}-2\mathbf{x}-\mathbf{w}+2\mathbf{y},\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
=&\langle\mathbf{z}^+-\mathbf{w}^+,\mathbf{z}-\mathbf{w}\rangle+\tfrac{\gamma^2}{\lambda}\langle\mathbf{s}^+-\mathbf{t}^+,(\mathbf{I}-\lambda\mathbf{A}\mathbf{A}^\top)(\mathbf{s}-\mathbf{t})\rangle\\
&+\langle\mathbf{z}-2\mathbf{x}+\tilde{\mathbf{z}}-\mathbf{w}+2\mathbf{y}-\tilde{\mathbf{w}},\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
=&-\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
&+\tfrac{1}{2}(\|\mathbf{z}^+-\mathbf{w}^+\|^2+\|\mathbf{z}-\mathbf{w}\|^2-\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2)\\
&+\tfrac{\gamma^2}{2\lambda}(\|\mathbf{s}^+-\mathbf{t}^+\|_{\mathbf{M}}^2+\|\mathbf{s}-\mathbf{t}\|_{\mathbf{M}}^2-\|\mathbf{s}^+-\mathbf{s}-\mathbf{t}^++\mathbf{t}\|_{\mathbf{M}}^2).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|(\mathbf{z}^+,\mathbf{s}^+)-(\mathbf{w}^+,\mathbf{t}^+)\|_{\mathbf{M}}^2\\
\leq&-2\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
&+\|(\mathbf{z},\mathbf{s})-(\mathbf{w},\mathbf{t})\|_{\mathbf{M}}^2-\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2-\tfrac{\gamma^2}{\lambda}\|\mathbf{s}^+-\mathbf{s}-\mathbf{t}^++\mathbf{t}\|_{\mathbf{M}}^2.
\end{aligned}\tag{11}$$

For the cross term, we have

$$\begin{aligned}
&-2\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{z}^+-\mathbf{z}+\mathbf{x}-\mathbf{w}^++\mathbf{w}-\mathbf{y}\rangle\\
=&-2\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\rangle-2\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle\\
\leq&-2\gamma\langle\nabla f(\mathbf{x})-\nabla f(\mathbf{y}),\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\rangle-2\gamma\beta\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|^2\\
\leq&\epsilon\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2+\left(\tfrac{\gamma^2}{\epsilon}-2\gamma\beta\right)\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|^2.
\end{aligned}$$

The first inequality comes from the cocoerciveness of ∇f, and the second inequality comes from Young's inequality. Therefore, if $\frac{\gamma^2}{\epsilon}-2\gamma\beta\leq 0$ and $\epsilon\leq 1$, we have

$$\begin{aligned}
&\|(\mathbf{z}^+,\mathbf{s}^+)-(\mathbf{w}^+,\mathbf{t}^+)\|_{\mathbf{M}}^2\\
\leq&\|(\mathbf{z},\mathbf{s})-(\mathbf{w},\mathbf{t})\|_{\mathbf{M}}^2+\left(\tfrac{\gamma^2}{\epsilon}-2\gamma\beta\right)\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|^2\\
&-(1-\epsilon)\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2-\tfrac{\gamma^2}{\lambda}\|\mathbf{s}^+-\mathbf{s}-\mathbf{t}^++\mathbf{t}\|_{\mathbf{M}}^2.
\end{aligned}\tag{12}$$

In addition, letting $\epsilon=\gamma/(2\beta)$, we have that

$$\begin{aligned}
&\|(\mathbf{z}^+,\mathbf{s}^+)-(\mathbf{w}^+,\mathbf{t}^+)\|_{\mathbf{M}}^2\\
\leq&\|(\mathbf{z},\mathbf{s})-(\mathbf{w},\mathbf{t})\|_{\mathbf{M}}^2-\tfrac{2\beta-\gamma}{2\beta}\left(\|(\mathbf{z}^+,\mathbf{s}^+)-(\mathbf{z},\mathbf{s})-(\mathbf{w}^+,\mathbf{t}^+)+(\mathbf{w},\mathbf{t})\|_{\mathbf{M}}^2\right).
\end{aligned}\tag{13}$$

that is, $\mathbf{T}$ is $\alpha-$averaged with $\alpha = \frac{1}{\frac{2\beta-\gamma}{2\beta}+1} = \frac{2\beta}{4\beta-\gamma}$ under the norm $\|(\cdot,\cdot)\|_{\mathbf{M}}$ for $\mathcal{X}\times\mathcal{Y}$.

Remark 1 For the three-operator splitting in [10] (i.e., $\mathbf{A}=\mathbf{I}$ and $\lambda=1$), we have $\mathbf{M}=\mathbf{0}$ and (12) becomes

$$\begin{aligned}\|\mathbf{z}^+-\mathbf{w}^+\|^2 \leq& \|\mathbf{z}-\mathbf{w}\|^2-(1-\epsilon)\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2\\ &+\left(\frac{\gamma^2}{\epsilon}-2\gamma\beta\right)\|\nabla f(\mathbf{x})-\nabla f(\mathbf{y})\|^2,\end{aligned}$$

which is Remark 3.1 in [10].

Remark 2 For the case with $f=0$ (Chambolle-Pock), we have, from (11),

$$\begin{aligned}&\|\mathbf{z}^+-\mathbf{w}^+\|^2+\frac{\gamma^2}{\lambda}\|\mathbf{s}^+-\mathbf{t}^+\|_{\mathbf{M}}^2\\ \leq&\|\mathbf{z}-\mathbf{w}\|^2+\tfrac{\gamma^2}{\lambda}\|\mathbf{s}-\mathbf{t}\|_{\mathbf{M}}^2-\|\mathbf{z}^+-\mathbf{z}-\mathbf{w}^++\mathbf{w}\|^2-\tfrac{\gamma^2}{\lambda}\|\mathbf{s}^+-\mathbf{s}-\mathbf{t}^++\mathbf{t}\|_{\mathbf{M}}^2,\end{aligned}$$

that is, the operator is firmly nonexpansive. This is shown in [14,21] by reformulating the Chambolle-Pock algorithm as a proximal point algorithm applied on the KKT conditions. Note that the previous result shows the nonexpansiveness of the Chambolle-Pock algorithm for $(\mathbf{x},\mathbf{s})$ under a norm defined by $\frac{1}{\gamma}\|\mathbf{x}\|^2-2\langle\mathbf{A}\mathbf{x},\mathbf{s}\rangle+\frac{\gamma}{\lambda}\|\mathbf{s}\|^2$ [21], while we show the non-expansiveness for $(\mathbf{z},\mathbf{s})$ under the norm defined by $\|(\cdot,\cdot)\|_{\mathbf{M}}$.

Remark 3 In [6], the PDFP^{2}O is shown to be nonexpansive only under a different norm for $\mathcal{X}\times\mathcal{Y}$ that is defined by $\sqrt{\|\mathbf{z}\|^2+\frac{\gamma^2}{\lambda}\|\mathbf{s}\|^2}$. Lemma 1 improves the result in [6] by showing that it is α-averaged.

Lemma 2 *For any fixed point $(\mathbf{z}^*,\mathbf{s}^*)$ of $\mathbf{T}$, $(\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z}^*)$ is an optimal solution to the optimization problem (1). For any optimal solution $\mathbf{x}^*$ of the optimization problem (1), we can find a fixed point $(\mathbf{z}^*,\mathbf{s}^*)$ of $\mathbf{T}$ such that $\mathbf{x}^*=(\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z}^*)$.*

Proof If $(\mathbf{z}^*,\mathbf{s}^*)$ is a fixed point of $\mathbf{T}$, let $\mathbf{x}^*=(\mathbf{I}+\gamma\partial g)^{-1}(\mathbf{z}^*)$. Then we have $\mathbf{0}=\mathbf{z}^*-\mathbf{x}^*+\gamma\nabla f(\mathbf{x}^*)+\gamma\mathbf{A}^\top\mathbf{s}^*$ from (2c), $\mathbf{z}^*-\mathbf{x}^*\in\gamma\partial g(\mathbf{x}^*)$ from (2a), and $\mathbf{A}\mathbf{x}^*\in\partial h^*(\mathbf{s}^*)$ from (2b) and (2c). Therefore, $\mathbf{0}\in\gamma\partial g(\mathbf{x}^*)+\gamma\nabla f(\mathbf{x}^*)+\gamma\mathbf{A}^\top\partial h(\mathbf{A}\mathbf{x}^*)$, i.e., $\mathbf{x}^*$ is an optimal solution for the convex problem (1).

If $\mathbf{x}^*$ is an optimal solution for problem (1), we have $\mathbf{0}\in\partial g(\mathbf{x}^*)+\nabla f(\mathbf{x}^*)+\mathbf{A}^\top\partial h(\mathbf{A}\mathbf{x}^*)$. Thus there exist $\mathbf{u}_g^*\in\partial g(\mathbf{x}^*)$ and $\mathbf{u}_h^*\in\partial h(\mathbf{A}\mathbf{x}^*)$ such that $\mathbf{0}=\mathbf{u}_g^*+\nabla f(\mathbf{x}^*)+\mathbf{A}^\top\mathbf{u}_h^*$. Letting $\mathbf{z}^*=\mathbf{x}^*+\gamma\mathbf{u}_g^*$ and $\mathbf{s}^*=\mathbf{u}_h^*$, we derive $\mathbf{x}=\mathbf{x}^*$ from (2a), $\mathbf{s}^+=(\mathbf{I}+\frac{\lambda}{\gamma}\partial h^*)^{-1}(\mathbf{s}^*+\frac{\lambda}{\gamma}\mathbf{A}\mathbf{x}^*)=\mathbf{s}^*$ from (2b), and $\mathbf{z}^+=\mathbf{x}^*-\gamma\nabla f(\mathbf{x}^*)-\gamma\mathbf{A}^\top\mathbf{s}^*=\mathbf{x}^*+\gamma\mathbf{u}_g^*=\mathbf{z}^*$ from (2c). Thus $(\mathbf{z}^*,\mathbf{s}^*)$ is a fixed point of $\mathbf{T}$.

Theorem 1 *1) Let $(\mathbf{z}^*,\mathbf{s}^*)$ be any fixed point of $\mathbf{T}$. Then $(\|(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}})_{k\geq0}$ is monotonically nonincreasing.*

2) The sequence $(\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}})_{k\geq 0}$ *is monotonically nonincreasing and converges to 0.*

3) We have the following convergence rate

$$\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2 \leq \frac{2\beta}{2\beta-\gamma}\frac{\|(\mathbf{z}^0,\mathbf{s}^0)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2}{k+1}$$

and

$$\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2 = o\left(\frac{1}{k+1}\right)$$

4) $(\mathbf{z}^k,\mathbf{s}^k)$ *weakly converges to a fixed point of* $\mathbf{T}$*, and if* $\mathcal{X}$ *has finite dimension, then it is strongly convergent.*

Proof 1) Let $(\mathbf{w},\mathbf{t})=(\mathbf{z}^*,\mathbf{s}^*)$ in (13), and we have that

$$\begin{aligned}&\|(\mathbf{z}^{k+1},\mathbf{s}^{k+1})-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2\\ \leq&\|(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2-\tfrac{2\beta-\gamma}{2\beta}\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2.\end{aligned}\tag{14}$$

Thus, $\|(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2$ is monotonically decreasing as long as $\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\neq\mathbf{0}$.

2) Summing up (14) from 0 to ∞, we have that

$$\textstyle\sum_{k=0}^{\infty}(\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2)\leq\frac{2\beta}{2\beta-\gamma}\|(\mathbf{z}^0,\mathbf{s}^0)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2$$

and thus the sequence $(\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}})_{k\geq 0}$ converges to 0. Furthermore, the sequence $(\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}})_{k\geq 0}$ is monotonically nonincreasing because $\|\mathbf{T}(\mathbf{z}^{k+1},\mathbf{s}^{k+1})-(\mathbf{z}^{k+1},\mathbf{s}^{k+1})\|_{\mathbf{M}}^2=\|\mathbf{T}(\mathbf{z}^{k+1},\mathbf{s}^{k+1})-\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2\leq\|(\mathbf{z}^{k+1},\mathbf{s}^{k+1})-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2=\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2$.

3) The convergence rate follows from [9, Lemma 3].

$$\begin{aligned}\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2\leq&\tfrac{1}{k+1}\textstyle\sum_{k=0}^{\infty}(\|\mathbf{T}(\mathbf{z}^k,\mathbf{s}^k)-(\mathbf{z}^k,\mathbf{s}^k)\|_{\mathbf{M}}^2)\\ \leq&\tfrac{2\beta}{2\beta-\gamma}\tfrac{\|(\mathbf{z}^0,\mathbf{s}^0)-(\mathbf{z}^*,\mathbf{s}^*)\|_{\mathbf{M}}^2}{k+1}.\end{aligned}$$

4) It follows from [1, Theorem 5.14].

Remark 4 Since the operator $\mathbf{T}$ is already α-averaged with $\alpha=\frac{2\beta}{4\beta-\gamma}$, we can enlarge the region of the relaxation parameter to $\theta_k\in(0,\frac{4\beta-\gamma}{2\beta})$, and the iteration $(\mathbf{z}^{k+1},\mathbf{s}^{k+1})=\mathbf{T}_{\theta_k}(\mathbf{z}^k,\mathbf{s}^k)$ still converges if $\sum_{k=0}^{\infty}\theta_k(\frac{4\beta-\gamma}{2\beta}-\theta_k)<\infty$.

3.3 Linear Convergence Rate for Special Cases

In this subsection, we provide some results on the linear convergence rate of PD3O with additional assumptions. For simplicity, we let $(\mathbf{z}^*, \mathbf{s}^*)$ be a fixed point of $\mathbf{T}$ and $\mathbf{x}^* \in (1+\gamma\partial g)^{-1}(\mathbf{z}^*)$. Then we denote

$$\mathbf{u}_h = \frac{\gamma}{\lambda}(\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s} + \mathbf{A}\tilde{\mathbf{z}} - \frac{\gamma}{\lambda}\mathbf{s}^+ \in \partial h^*(\mathbf{s}^+), \tag{15}$$

$$\mathbf{u}_g = \frac{1}{\gamma}(\mathbf{z} - \mathbf{x}) \in \partial g(\mathbf{x}), \tag{16}$$

$$\mathbf{u}_h^* = \mathbf{A}(\tilde{\mathbf{z}}^* - \gamma\mathbf{A}^\top\mathbf{s}^*) = \mathbf{A}\mathbf{x}^* \in \partial h^*(\mathbf{s}^*), \tag{17}$$

$$\mathbf{u}_g^* = \frac{1}{\gamma}(\mathbf{z}^* - \mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{18}$$

In addition, we let $\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{u}_h - \mathbf{u}_h^* \rangle \geq \tau_h \|\mathbf{s}^+ - \mathbf{s}^*\|_{\mathbf{M}}^2$ for any $\mathbf{s}^+ \in \mathcal{Y}$ and $\mathbf{u}_h \in \partial h^*(\mathbf{s}^+)$. Then ∂h^* is τ_h-strongly monotone under the norm defined by $\mathbf{M}$. Here we allow that $\tau_h = 0$ for just monotone operators. Similarly, we let $\langle \mathbf{x} - \mathbf{x}^*, \mathbf{u}_g - \mathbf{u}_g^* \rangle \geq \tau_g \|\mathbf{x} - \mathbf{x}^*\|^2$ and $\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*) \rangle \geq \tau_f \|\mathbf{x} - \mathbf{x}^*\|$ for any $\mathbf{x} \in \mathcal{X}$.

Theorem 2 *If g has a L_g-Lipschitz continuous gradient, i.e.,*

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L_g \|\mathbf{x} - \mathbf{y}\|,$$

then we have

$$\|\mathbf{z}^+ - \mathbf{z}^*\|^2 + \left(\frac{\gamma^2}{\lambda} + \tau_h\right)\|\mathbf{s}^+ - \mathbf{s}^*\|_{\mathbf{M}}^2 \leq \rho\left(\|\mathbf{z} - \mathbf{z}^*\|^2 + \left(\frac{\gamma^2}{\lambda} + \tau_h\right)\|\mathbf{s} - \mathbf{s}^*\|_{\mathbf{M}}^2\right)$$

where

$$\rho = \max\left(\frac{\gamma}{\gamma + 2\tau_h\lambda}, 1 - \frac{\left(2\gamma - \frac{\gamma^2}{\beta}\right)\tau_f + 2\gamma\tau_g}{1 + \gamma L_g}\right). \tag{19}$$

When, in addition, $\gamma < 2\beta$, $\tau_h > 0$, and $\tau_f + \tau_g > 0$, we have that $\rho < 1$ and the algorithm converges linearly.

Proof First of all, we have the equalities (20) and (21):

$$\begin{aligned}
&\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{u}_h - \mathbf{u}_h^* \rangle \\
=&\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \tfrac{\gamma}{\lambda}(\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s} + \mathbf{A}\tilde{\mathbf{z}} - \tfrac{\gamma}{\lambda}\mathbf{s}^+ - \mathbf{A}\mathbf{x}^* \rangle \\
=&\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \tfrac{\gamma}{\lambda}(\mathbf{I} - \lambda\mathbf{A}\mathbf{A}^\top)\mathbf{s} + \mathbf{A}(\mathbf{z}^+ - \mathbf{z} + \mathbf{x} + \gamma\mathbf{A}^\top\mathbf{s}^+) - \tfrac{\gamma}{\lambda}\mathbf{s}^+ - \mathbf{A}\mathbf{x}^* \rangle \\
=&\tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{s} - \mathbf{s}^+ \rangle_{\mathbf{M}} + \gamma\langle \mathbf{A}^\top\mathbf{s}^+ - \mathbf{A}^\top\mathbf{s}^*, \mathbf{z}^+ - \mathbf{z} + \mathbf{x} - \mathbf{x}^* \rangle,
\end{aligned} \tag{20}$$

where the first equality comes from the definitions of $\mathbf{u}_h$ (15) and $\mathbf{u}_h^*$ (17) and the second equality comes from the definition of $\tilde{\mathbf{z}}$ (9).

$$\begin{aligned}
&\gamma\langle \mathbf{x} - \mathbf{x}^*, \mathbf{u}_g - \mathbf{u}_g^* \rangle + \gamma\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*) \rangle \\
=&\gamma\langle \mathbf{x} - \mathbf{x}^*, \tfrac{1}{\gamma}(\mathbf{z} - \mathbf{x}) - \tfrac{1}{\gamma}(\mathbf{z}^* - \mathbf{x}^*) \rangle + \gamma\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*) \rangle \\
=&\langle \mathbf{x} - \mathbf{x}^*, \mathbf{z} - \mathbf{x} - \mathbf{z}^* + \mathbf{x}^* + \gamma\nabla f(\mathbf{x}) - \gamma\nabla f(\mathbf{x}^*) \rangle \\
=&\langle \mathbf{x} - \mathbf{x}^*, \mathbf{z} - \mathbf{z}^+ - \gamma\mathbf{A}^\top\mathbf{s}^+ + \gamma\mathbf{A}^\top\mathbf{s}^* \rangle,
\end{aligned} \tag{21}$$

where the first equality comes from the definitions of $\mathbf{u}_g$ (16) and $\mathbf{u}_g^*$ (18) and the last equality comes from the update of $\mathbf{z}^+$ in (2c).

Combing both (20) and (21), we have

$$\begin{aligned}
&\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{u}_h - \mathbf{u}_h^*\rangle + \gamma\langle \mathbf{x} - \mathbf{x}^*, \mathbf{u}_g - \mathbf{u}_g^*\rangle + \gamma\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\rangle\\
=&\tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{s} - \mathbf{s}^+\rangle_{\mathbf{M}} + \gamma\langle \mathbf{A}^\top\mathbf{s}^+ - \mathbf{A}^\top\mathbf{s}^*, \mathbf{z}^+ - \mathbf{z} + \mathbf{x} - \mathbf{x}^*\rangle\\
&+ \langle \mathbf{x} - \mathbf{x}^*, \mathbf{z} - \mathbf{z}^+ - \gamma\mathbf{A}^\top\mathbf{s}^+ + \gamma\mathbf{A}^\top\mathbf{s}^*\rangle\\
=&\tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{s} - \mathbf{s}^+\rangle_{\mathbf{M}} + \langle \mathbf{x} - \gamma\mathbf{A}^\top\mathbf{s}^+ - \mathbf{x}^* + \gamma\mathbf{A}^\top\mathbf{s}^*, \mathbf{z} - \mathbf{z}^+\rangle\\
=&\tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{s} - \mathbf{s}^+\rangle_{\mathbf{M}} + \langle \mathbf{z}^+ + \gamma\nabla f(\mathbf{x}) - \mathbf{z}^* - \gamma\nabla f(\mathbf{x}^*), \mathbf{z} - \mathbf{z}^+\rangle\\
=&\tfrac{\gamma^2}{\lambda}\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{s} - \mathbf{s}^+\rangle_{\mathbf{M}} + \langle \mathbf{z}^+ - \mathbf{z}^*, \mathbf{z} - \mathbf{z}^+\rangle + \gamma\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*), \mathbf{z} - \mathbf{z}^+\rangle\\
=&\tfrac{1}{2}\|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2 - \tfrac{1}{2}\|(\mathbf{z}^+, \mathbf{s}^+) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2 - \tfrac{1}{2}\|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^+, \mathbf{s}^+)\|_{\mathbf{M}}^2\\
&- \gamma\langle \mathbf{z} - \mathbf{z}^+, \nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x})\rangle,
\end{aligned}$$

where the third equality comes from the update of $\mathbf{z}^+$ in (2c) and the last equality comes from $2\langle a, b\rangle = \|a + b\|^2 - \|a\|^2 - \|b\|^2$. Rearranging it, we have

$$\begin{aligned}
&\|(\mathbf{z}^+, \mathbf{s}^+) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2\\
=&\|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2 - \|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^+, \mathbf{s}^+)\|_{\mathbf{M}}^2 - 2\gamma\langle \mathbf{z} - \mathbf{z}^+, \nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x})\rangle\\
&- 2\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{u}_h - \mathbf{u}_h^*\rangle - 2\gamma\langle \mathbf{x} - \mathbf{x}^*, \mathbf{u}_g - \mathbf{u}_g^*\rangle - 2\gamma\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\rangle.
\end{aligned}$$

The Young's inequality gives us

$$-2\gamma\langle \mathbf{z} - \mathbf{z}^+, \nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x})\rangle \le \|\mathbf{z} - \mathbf{z}^+\|^2 + \gamma^2\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2,$$

and the cocoerciveness of ∇f shows

$$-\tfrac{\gamma^2}{\beta}\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\rangle \le -\gamma^2\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2.$$

Therefore, we have

$$\begin{aligned}
&\|(\mathbf{z}^+, \mathbf{s}^+) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2\\
\le&\|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2 - \tfrac{\gamma^2}{\lambda}\|\mathbf{s} - \mathbf{s}^+\|_{\mathbf{M}}^2 - \left(2\gamma - \tfrac{\gamma^2}{\beta}\right)\langle \mathbf{x} - \mathbf{x}^*, \nabla f(\mathbf{x}) - \nabla f(\mathbf{x}^*)\rangle\\
&- 2\gamma\langle \mathbf{s}^+ - \mathbf{s}^*, \mathbf{u}_h - \mathbf{u}_h^*\rangle - 2\gamma\langle \mathbf{x} - \mathbf{x}^*, \mathbf{u}_g - \mathbf{u}_g^*\rangle\\
\le&\|(\mathbf{z}, \mathbf{s}) - (\mathbf{z}^*, \mathbf{s}^*)\|_{\mathbf{M}}^2 - \left(\left(2\gamma - \tfrac{\gamma^2}{\beta}\right)\tau_f + 2\gamma\tau_g\right)\|\mathbf{x} - \mathbf{x}^*\|^2 - 2\gamma\tau_h\|\mathbf{s}^+ - \mathbf{s}^*\|_{\mathbf{M}}^2.
\end{aligned}$$

Thus

$$\begin{aligned}
&\|\mathbf{z}^+ - \mathbf{z}^*\|^2 + \left(\tfrac{\gamma^2}{\lambda} + 2\gamma\tau_h\right)\|\mathbf{s}^+ - \mathbf{s}^*\|_{\mathbf{M}}^2\\
\le&\|\mathbf{z} - \mathbf{z}^*\|^2 + \tfrac{\gamma^2}{\lambda}\|\mathbf{s} - \mathbf{s}^*\|_{\mathbf{M}}^2 - \left(\left(2\gamma - \tfrac{\gamma^2}{\beta}\right)\tau_f + 2\gamma\tau_g\right)\|\mathbf{x} - \mathbf{x}^*\|^2\\
\le&\|\mathbf{z} - \mathbf{z}^*\|^2 + \tfrac{\gamma^2}{\lambda}\|\mathbf{s} - \mathbf{s}^*\|_{\mathbf{M}}^2 - \tfrac{\left(\left(2\gamma - \frac{\gamma^2}{\beta}\right)\tau_f + 2\gamma\tau_g\right)}{1 + \gamma L_g}\|\mathbf{z} - \mathbf{z}^*\|^2\\
\le&\rho\left(\|\mathbf{z} - \mathbf{z}^*\|^2 + \left(\tfrac{\gamma^2}{\lambda} + 2\gamma\tau_h\right)\|\mathbf{s} - \mathbf{s}^*\|_{\mathbf{M}}^2\right),
\end{aligned}$$

with ρ defined in (19).

4 Numerical Experiments

In this section, we will compare the proposed algorithm PD3O with PDFP and the Condat-Vu algorithm in solving two problems: the fused lasso problem and an image reconstruction problem. In fact, PDFP can be reformulated in the same way as (3) for PD3O and (8) for Condat-Vu by letting $\mathbf{z}$ be $\bar{\mathbf{x}}$ and shifting the updating order to $(\mathbf{s},\ \mathbf{x},\ \bar{\mathbf{x}})$. Its reformulation is

$$\mathbf{s}^+ = (\mathbf{I} + \tfrac{\lambda}{\gamma}\partial h^*)^{-1}\left(\mathbf{s} + \tfrac{\lambda}{\gamma}\mathbf{A}\bar{\mathbf{x}}\right); \tag{22a}$$

$$\mathbf{x}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x} - \gamma\nabla f(\mathbf{x}) - \gamma\mathbf{A}^\top\mathbf{s}^+); \tag{22b}$$

$$\bar{\mathbf{x}}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x}^+ - \gamma\nabla f(\mathbf{x}^+) - \gamma\mathbf{A}^\top\mathbf{s}^+). \tag{22c}$$

Therefore, the difference between these three algorithms is in the third step for updating $\bar{\mathbf{x}}$. The third steps for these three algorithms are summarized below:

PDFP	$\bar{\mathbf{x}}^+ = (\mathbf{I} + \gamma\partial g)^{-1}(\mathbf{x}^+ - \gamma\nabla f(\mathbf{x}^+) - \gamma\mathbf{A}^\top\mathbf{s}^+)$
Condat-Vu	$\bar{\mathbf{x}}^+ = 2\mathbf{x}^+ - \mathbf{x}$
PD3O	$\bar{\mathbf{x}}^+ = 2\mathbf{x}^+ - \mathbf{x} + \gamma\nabla f(\mathbf{x}) - \gamma\nabla f(\mathbf{x}^+)$

Though there are two more terms ($\nabla f(\mathbf{x})$ and $\nabla f(\mathbf{x}^+)$) in PD3O than the Condat-Vu algorithm, $\nabla f(\mathbf{x})$ has been computed in the previous step, and $\nabla f(\mathbf{x}^+)$ will be used in the next iteration. Thus, except that $\nabla f(\mathbf{x})$ has to be stored, there is no additional cost in PD3O comparing to the Condat-Vu algorithm. However, for PDFP, the resolvent $(\mathbf{I} + \gamma\partial g)^{-1}$ is applied twice on different values in each iteration, and it will not be used in the next iteration. Therefore, the per-iteration cost is more than the other two algorithms. When this resolvent is easy to compute or has a closed-form solution such that its cost is much smaller than other operators, we can ignore the additional cost. So we compare the number of iterations only in the numerical experiments.

4.1 Fused lasso problem

The fused lasso, which was proposed in [23], has been applied in many fields. The problem with the least squares loss is

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \mu_1\|\mathbf{x}\|_1 + \mu_2\sum_{i=1}^{p-1}|x_{i+1} - x_i|,$$

where $\mathbf{x} = (x_1, \cdots, x_p) \in \mathbf{R}^p$, $\mathbf{A} \in \mathbf{R}^{n\times p}$, and $\mathbf{b} \in \mathbf{R}^n$. We use the same setting as [7] for the numerical experiments. Let $n = 500$ and $p = 10,000$. The true sparse vector $\mathbf{x}$ is shown in Fig. 1. The matrix $\mathbf{A}$ is a random matrix whose elements follow the standard Gaussian distribution, and $\mathbf{b}$ is obtained by adding independent and identically distributed (i.i.d.) Gaussian noise with variance 0.01 onto $\mathbf{A}\mathbf{x}$. For the penalty parameters, we set $\mu_1 = 20$ and $\mu_2 = 200$.

We would like to compare the three algorithms with different parameters and the results can guide us in choosing parameters for these algorithms in

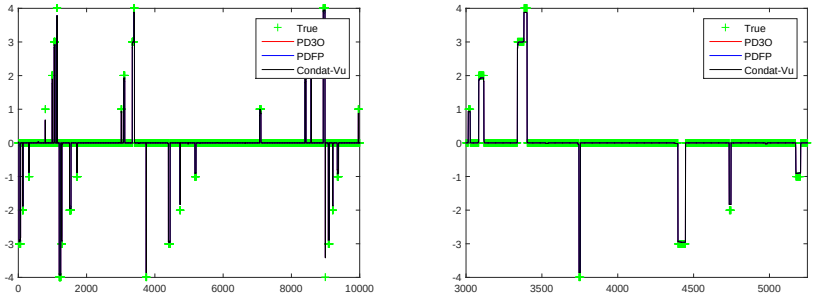

Fig. 1: The true sparse signal and the reconstructed results using PD3O, PDFP, and Condat-Vu. The right figure is a zoom-in of the signal in [3000, 5500]. All these three algorithms have the same result.

other applications. Recall that the parameters for the Condat-Vu algorithm have to satisfy $\lambda\left(2-2\cos(\frac{p-1}{p}\pi)\right)+\gamma/(2\beta)\leq 1$, and those for PD3O and PDFP have to satisfy $\lambda\left(2-2\cos(\frac{p-1}{p}\pi)\right)<1$ and $\gamma<2\beta$. Firstly, we fix $\lambda=1/8$ and let $\gamma=\beta,\ 1.5\beta,\ 1.9\beta$. The objective values for these algorithms after each iteration are compared in Fig. 2 (Left). The approximate optimal objective value f^* is obtained by running PD3O for 10,000 iterations. The results show that the three algorithms have very close performance when they converge ($\gamma=\beta,\ 1.5\beta$[2]) and PD3O is slightly better than PDFP for a large γ. However, they converge faster with a larger stepsize γ if they converge. Therefore, having a large range for acceptable parameters ensuring convergence is important.

Then we fix $\gamma=1.9\beta$ and let $\lambda=1/80,\ 1/8,\ 1/4$. The objective values for these algorithms after each iteration are compared in Fig. 2 (Right). Again, we can see that the performances for these three algorithms are very close when they converge, and PD3O is slightly better than PDFP for a large λ at later iterations. This result also suggests that it is better to choose a slightly large λ ($\lambda\geq\beta$) and the increase in λ does not bring too much advantage in the first several iterations if λ is large enough. Both experiments demonstrate the effectiveness of having a large range of acceptable parameters.

[2] Condat-Vu is not guaranteed to converge when $\gamma=1.5\beta$ though it converges in this case. We tried a non-zero initial $\mathbf{x}$ and it diverges.

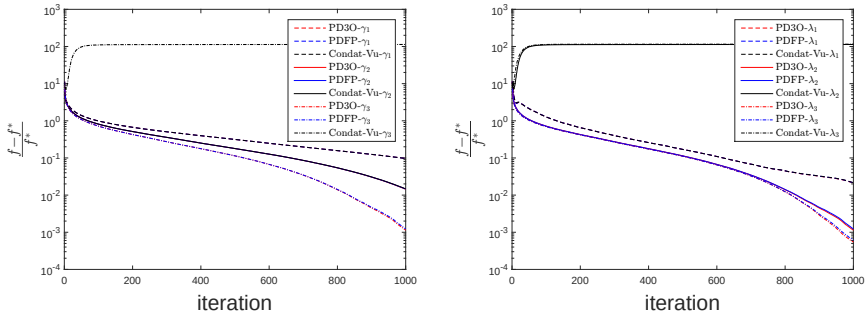

Fig. 2: The comparison of these algorithms on the fused lasso problem. In the left figure, we fix $\lambda = 1/8$ and let $\gamma = \beta,\ 1.5\beta,\ 1.9\beta$. In the right figure, we fix $\gamma = 1.9\beta$ and let $\lambda = 1/80,\ 1/8,\ 1/4$. PD3O and PDFP perform better than the Condat-Vu algorithm because they have a larger range for acceptable parameters and choosing large numbers for both parameters makes the algorithm converge fast. In addition, FD3O performs slightly better than PDFP.

4.2 Computed tomography reconstruction problem

In this subsection, we compare the three algorithms on Computed Tomography (CT) image reconstruction. The optimization problem can be formulated as

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \mu\|\mathbf{D}\mathbf{x}\|_1 + \iota_C(\mathbf{x}),$$

where $\mathbf{x} \in \mathbf{R}^n$ is the image to be reconstructed (n is the number of pixels), $\mathbf{A} \in \mathbf{R}^{m\times n}$ is the forward projection matrix that maps the image to the sinogram data, $\mathbf{b} \in \mathbf{R}^m$ is the measured noisy sinogram data, $\mathbf{D}$ is a discrete gradient operator, and ι_C is the indicator function that returns zero if $x_i \geq 0$ for all $1 \leq i \leq n$ and $+\infty$ otherwise. We use the same dataset as [26] in the numerical experiments. The true image to be recovered is the 128×128 Shepp-Logan phantom image ($n = 128 \times 128 = 16,384$), and the sinogram data is measured from 50 uniformly oriented projections ($m = 9,250$) with white Gaussian noise of mean 0 and variance 1. For the penalty parameter, we let $\mu = 0.05$. Similarly, we compare the performance of these algorithm for different parameters and the result is show in Fig. 3.

Firstly, we fix $\lambda = 1/16$ and let $\gamma = \beta,\ 1.5\beta,\ 1.9\beta$. The objective values for these algorithms after each iteration are compared in Fig. 3 (Left). The approximate optimal objective value f^* is obtained by running PD3O for 10,000 iterations. The results show that the three algorithms have very close performance when they converge[3]. However, they converge faster with a larger stepsize γ if they converge. Then we fix $\gamma = 1.9\beta$ and let $\lambda = 1/160,\ 1/16,\ 1/8$.

[3] Condat-Vu is not guaranteed to converge when $\gamma = 1.5\beta,\ 1.9\beta$.

The objective values for these algorithms after each iteration are compared in Fig. 3 (Right). Again, we can see that the performances for these three algorithms are very close when they converge.

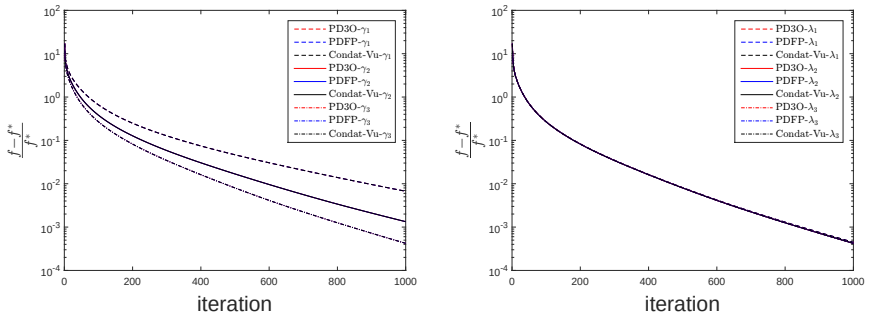

Fig. 3: The comparison of these algorithms on the CT reconstruction problem. In the left figure, we fix $\lambda = 1/16$ and let $\gamma = \beta,\ 1.5\beta,\ 1.9\beta$. In the right figure, we fix $\gamma = 1.9\beta$ and let $\lambda = 1/160,\ 1/16,\ 1/8$. PD3O and PDFP perform better than the Condat-Vu algorithm because they have a larger range for acceptable parameters and choosing large numbers for both parameters makes the algorithm converge fast. In addition, FD3O performs slightly better than PDFP.

5 Conclusion

In this paper, we proposed a novel primal-dual three-operator splitting scheme PD3O for solving $f(\mathbf{x}) + g(\mathbf{x}) + h(\mathbf{A}\mathbf{x})$. It has the famous primal-dual algorithms PDHG and PDFP^{2}O for solving the sum of two functions as special cases. Comparing to the two existing primal-dual algorithms PDFP and the Condat-Vu algorithm for solving the sum of three functions, PD3O has the advantages from both algorithms–a low per-iteration complexity and a large range of acceptable parameters ensuring the convergence. The numerical experiments show the effectiveness and efficiency of PD3O. We left the acceleration of PD3O as the future work.

Acknowledgment

The author would like thank Dr. Peijun Chen for providing the code for PDFP.

References

1. H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, Springer Science & Business Media, 2011. 3.1, 3.2
2. I. M. Baytas, M. Yan, A. K. Jain, and J. Zhou, *Asynchronous multi-task learning*, in IEEE International Conference on Data Mining (ICDM), IEEE, 2016, p. to appear. 1
3. L. M. Briceo-Arias, *Forward-Douglas-Rachford splitting and forward-partial inverse method for solving monotone inclusions*, Optimization, 64 (2015), pp. 1239–1261. 1
4. A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145. 1, 1, 2, 2.2
5. ———, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319. 1
6. P. Chen, J. Huang, and X. Zhang, *A primal–dual fixed point algorithm for convex separable minimization with applications to image restoration*, Inverse Problems, 29 (2013), p. 025011. 1, 1, 2, 2.1, 3
7. ———, *A primal-dual fixed point algorithm for minimization of the sum of three convex separable functions*, Fixed Point Theory and Applications, 2016 (2016), pp. 1–18. 1, 1, 2, 2.4, 4.1
8. L. Condat, *A primal–dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms*, Journal of Optimization Theory and Applications, 158 (2013), pp. 460–479. 1, 1, 2, 2.5
9. D. Davis and W. Yin, *Convergence rate analysis of several splitting schemes*, arXiv preprint arXiv:1406.4834, (2014). 3.2
10. ———, *A three-operator splitting scheme and its optimization applications*, arXiv preprint arXiv:1504.01032, (2015). 1, 1, 2, 2.3, 1
11. J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transaction of the American Mathematical Society, 82 (1956), pp. 421–489. 1
12. E. Esser, X. Zhang, and T. F. Chan, *A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046. 1
13. D. Gabay, *Applications of the method of multipliers to variational inequalities*, Studies in Mathematics and its Applications, 15 (1983), pp. 299–331. 1
14. B. He, Y. You, and X. Yuan, *On the convergence of primal-dual hybrid gradient algorithm*, SIAM Journal on Imaging Sciences, 7 (2014), pp. 2526–2537. 2
15. N. Komodakis and J. C. Pesquet, *Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems*, IEEE Signal Processing Magazine, 32 (2015), pp. 31–54. 1
16. A. Krol, S. Li, L. Shen, and Y. Xu, *Preconditioned alternating projection algorithms for maximum a posteriori ECT reconstruction*, Inverse problems, 28 (2012), p. 115005. 2.4
17. P. Latafat and P. Patrinos, *Asymmetric forward-backward-adjoint splitting for solving monotone inclusions involving three operators*, arXiv preprint arXiv:1602.08729, (2016). 1
18. I. Loris and C. Verhoeven, *On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty*, Inverse Problems, 27 (2011), p. 125007. 1, 2, 2.1
19. G. B. Passty, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, Journal of Mathematical Analysis and Applications, 72 (1979), pp. 383 – 390. 1
20. D. W. Peaceman and J. H. H. Rachford, *The numerical solution of parabolic and elliptic differential equations*, Journal of the Society for Industrial and Applied Mathematics, 3 (1955), pp. 28–41. 1
21. Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin, *Coordinate friendly structures, algorithms and applications*, Annals of Mathematical Sciences and Applications, 1 (2016), pp. 57–119. 1, 2

22. T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, in 2009 IEEE 12th International Conference on Computer Vision, IEEE, 2009, pp. 1133–1140. 1
23. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, *Sparsity and smoothness via the fused lasso*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 91–108. 1, 4.1
24. B. C. Vũ, *A splitting algorithm for dual monotone inclusions involving cocoercive operators*, Advances in Computational Mathematics, 38 (2013), pp. 667–681. 1, 1, 2, 2.5
25. M. Yan and W. Yin, *Self equivalence of the alternating direction method of multipliers*, arXiv preprint arXiv:1407.7400, (2014). 1
26. X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, Journal of Scientific Computing, 46 (2011), pp. 20–46. 4.2
27. H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67 (2005), pp. 301–320. 1