Journal of Optimization Theory and Applications (2022) 195:102–130
https://doi.org/10.1007/s10957-022-02061-8

Dualize, Split, Randomize: Toward Fast Nonsmooth Optimization Algorithms

Adil Salim[1] · Laurent Condat[1] · Konstantin Mishchenko[1] · Peter Richtárik[1]

Received: 19 July 2021 / Accepted: 5 June 2022 / Published online: 13 July 2022

Abstract

We consider minimizing the sum of three convex functions, where the first one F is smooth, the second one is nonsmooth and proximable and the third one is the composition of a nonsmooth proximable function with a linear operator L. This template problem has many applications, for instance, in image processing and machine learning. First, we propose a new primal–dual algorithm, which we call PDDY, for this problem. It is constructed by applying Davis–Yin splitting to a monotone inclusion in a primal–dual product space, where the operators are monotone under a specific metric depending on L. We show that three existing algorithms (the two forms of the Condat–Vũ algorithm and the PD3O algorithm) have the same structure, so that PDDY is the fourth missing link in this self-consistent class of primal–dual algorithms. This representation eases the convergence analysis: it allows us to derive sublinear convergence rates in general, and linear convergence results in presence of strong convexity. Moreover, within our broad and flexible analysis framework, we propose new stochastic generalizations of the algorithms, in which a variance-reduced random estimate of the gradient of F is used, instead of the true gradient. Furthermore, we obtain, as a special case of PDDY, a linearly converging algorithm for the minimization of a strongly

Communicated by Shoham Sabach.

✉ Laurent Condat
laurent.condat@kaust.edu.sa

Adil Salim
Adil.Salim@kaust.edu.sa

Konstantin Mishchenko
Konstantin.Mishchenko@kaust.edu.sa

Peter Richtárik
Peter.richtarik@kaust.edu.sa

[1] Computer Science Program, Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

convex function F under a linear constraint; we discuss its important application to decentralized optimization.

1 Introduction

Many problems in statistics, machine learning or signal processing can be formulated as high-dimensional convex optimization problems [3, 12, 56, 59, 68, 69]. They typically involve a smooth term F and a nonsmooth regularization term G, and $F + G$ is often minimized using (a variant of) Stochastic Gradient Descent (SGD) [47]. However, in many cases, G is not proximable; that is, its proximity operator does not admit a closed-form expression. In particular, structured regularization, like the total variation or its variants for images or graphs [10, 23, 24, 27, 31, 61, 63], or the overlapping group lasso [3], is known to have computationally expensive proximity operators. Also, when G is a sum of several regularizers, G is not proximable, even if the individual regularizers are, in general [60]. Thus, in many situations, G is not proximable, but it takes the form $G = R + H \circ L$ where R, H are proximable and L is a linear operator. Therefore, in this paper, we study the problem

$$\textbf{Problem (1)} : \underset{x \in \mathcal{X}}{\text{minimize}} \ \Big(F(x) + R(x) + H(Lx)\Big), \tag{1}$$

where $L : \mathcal{X} \to \mathcal{Y}$ is a linear operator, $\mathcal{X}$ and $\mathcal{Y}$ are real Hilbert spaces (all spaces are supposed of finite dimension), $F : \mathcal{X} \to \mathbb{R}$ is a convex function, $R : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ and $H : \mathcal{Y} \to \mathbb{R} \cup \{+\infty\}$ are proper, convex, lower semicontinuous functions; we refer to textbooks like [5, 9] for standard definitions of convex analysis. F is supposed to be ν-smooth, for some $\nu > 0$; that is, it is differentiable on $\mathcal{X}$ and its gradient ∇F is ν-Lipschitz continuous: $\|\nabla F(x) - \nabla F(x')\| \leq \nu \|x - x'\|$, for every $(x, x') \in \mathcal{X}^2$.

Our contributions are the following. We recast Problem (1) as finding a zero of the sum of three operators, which are monotone in a primal–dual product space, under a particular metric (Sect. 2). Then, we apply Davis–Yin splitting (DYS) [28], a generic method for this type of monotone inclusions (Sect. 3). By doing so, we recover the existing PD3O [77] and two forms of the Condat–Vũ [22, 72] algorithms, but we also discover a new one, which we call the Primal–Dual Davis–Yin (PDDY) algorithm (Sect. 4). In other words, we discover PDDY as the fourth "missing link" in a group of primal–dual algorithms, which is self-consistent, in the sense that by exchanging the roles of the primal and dual terms $R + H \circ L$ and $R^* \circ (-L^*) + H^*$, or by exchanging the roles of two monotone operators in the construction, we recover this or that algorithm. Furthermore, the decomposition of the primal–dual monotone inclusion into three terms allows us to use an important inequality regarding DYS for the analysis of the algorithms. More precisely, we can apply Lemma 3.2, by instantiating the monotone operators and inner product with the ones at hand. Thanks to this property, we can easily replace the gradient ∇F by a stochastic variance-reduced (VR) estimator, which can be much cheaper to evaluate (Sect. 5). Thus, we derive the first VR stochastic algorithms to tackle Problem (1), to the best of our knowledge. We also leverage the DYS representation of the algorithms to prove convergence rates; our analysis covers the deterministic and stochastic cases in a unified way (Sect. 5). Moreover, as a byproduct of our analysis, we discover the first linearly converging algorithm for the minimization of a smooth strongly convex function, using its gradient, under a linear constraint, without projections on it (Sect. 6). Its application to decentralized optimization is discussed in Appendix 1. Finally, numerical experiments illustrating the performance of the algorithms are presented in Sect. 7. A part of the proofs is deferred to Appendix 1, and additional linear convergence results are derived in Appendix 1.

1.1 Related Work

Splitting Algorithms: Algorithms allowing to solve nonsmooth optimization problem involving several proximable terms are called proximal splitting algorithms [6, 7, 19, 21, 25, 44, 57]. A classical one is the Douglas–Rachford algorithm [32, 51, 62, 70] (or, equivalently, the ADMM [8, 34, 35]) to minimize the sum of two nonsmooth functions $R+H$. To minimize $G = R+H \circ L$, the Douglas–Rachford algorithm can be generalized to the Primal–Dual Hybrid Gradient (PDHG) algorithm, a.k.a. Chambolle–Pock algorithm [11, 55]. Behind its success is the ability to handle the composite term $H \circ L$ using separate activation of L, its adjoint operator L^T, and the proximity operator of H. However, in many applications, the objective function involves a smooth function F, for instance, a least-squares term or a sum of logistic losses composed with inner products. To cover these applications, proximal splitting algorithms like the Combettes–Pesquet [20], Condat–Vũ [22, 72] and PD3O [77] algorithms have been proposed; they can solve the general Problem (1). These algorithms are primal–dual in nature; that is, they solve not only the primal problem (1), but also the dual problem, in a joint way. Many other algorithms exist to solve Problem (1), and we refer to [25] and [21] for an overview of primal–dual proximal splitting algorithms. We can also mention the class of *projective splitting* algorithms first proposed in [33] and further developed in several papers [2, 17, 41–43]. They proceed by building a separating hyperplane between the current iterate and the solution and then projecting the current iterate onto this hyperplane, to get closer to the solution. The projective splitting algorithms with forward steps [42, 43] are fully split and can solve Problem (1), as well.

Let us turn our attention to stochastic splitting algorithms. In machine learning applications, the gradient of F is often much too expensive to evaluate and replaced by a cheaper stochastic estimate. We can distinguish the two classes of standard stochastic gradients [36, 37, 47] and variance-reduced (VR) stochastic gradients [29, 36, 38, 40, 74, 79]. VR stochastic gradients are estimators that ensure convergence to an exact solution of the problem, like with deterministic algorithms; that is, the variance of the stochastic errors they induce tends to zero. For some problems, VR stochastic algorithms are significantly faster than their deterministic counterparts. By contrast, with standard stochastic gradients and constant stepsizes, the algorithms typically do not converge to a fixed point and continue to fluctuate in a neighborhood of the solution set; this can be sufficient if the desired accuracy is low and speed is critical. When $L = I$, where I denotes the identity, solving Problem (1) with standard and with VR stochastic gradients was considered in [78] and in [58], respectively. In the general case $L \neq I$ of interest in this paper, solving the problem with a standard stochastic gradient was considered in [80]. Thus, our proposed method is the first to allow solving the general Problem (1) in a flexible way, with calls to ∇F or to standard or VR stochastic estimates.

1.2 Mathematical Background

We introduce some notions and notations of convex analysis and operator theory, see [5, 9] for more details. Let $\mathcal{Z}$ be a real Hilbert space. Let $G : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The domain of G is the convex set $\mathrm{dom}(G) = \{z \in \mathcal{Z} \ : \ G(z) \neq +\infty\}$, its subdifferential is the set-valued operator $\partial G : z \in \mathcal{Z} \mapsto \{y \in \mathcal{Z} \ : \ (\forall z' \in \mathcal{Z})\ G(z) + \langle z' - z, y\rangle \leq G(z')\}$, and its conjugate function is $G^* : z \mapsto \sup_{z' \in \mathcal{Z}} \{\langle z, z'\rangle - G(z')\}$. If G is differentiable at $z \in \mathcal{Z}$, $\partial G(z) = \{\nabla G(z)\}$. We define the proximity operator of G as the operator $\mathrm{prox}_G : z \in \mathcal{Z} \mapsto \arg\min_{z' \in \mathcal{Z}} \left\{G(z') + \frac{1}{2}\|z - z'\|^2\right\}$. Finally, given any $b \in \mathcal{Z}$, we define the convex indicator function $\iota_b : z \mapsto \{0$ if $z = b$, $+\infty$ otherwise$\}$.

Let $M : \mathcal{Z} \to 2^{\mathcal{Z}}$ be a set-valued operator. The inverse M^{-1} of M is defined by the relation $z' \in M(z) \Leftrightarrow z \in M^{-1}(z')$. The set of zeros of M is $\mathrm{zer}(M) := \{z \in \mathcal{Z}, 0 \in M(z)\}$. M is monotone if $\langle w - w', z - z'\rangle \geq 0$ and strongly monotone if there exists $\mu > 0$ such that $\langle w - w', z - z'\rangle \geq \mu\|z - z'\|^2$, for every $(x, x') \in \mathcal{Z}^2$, $w \in M(z)$, $w' \in M(z')$. M is maximally monotone if its graph is not contained in the graph of another monotone operator. The resolvent of M is $J_M := (I + M)^{-1}$. If G is proper, convex and lower semicontinuous, ∂G is maximally monotone, $J_{\partial G} = \mathrm{prox}_G$, $\mathrm{zer}(\partial G) = \arg\min G$ and $(\partial G)^{-1} = \partial G^*$.

A single-valued operator M on $\mathcal{Z}$ is ξ-cocoercive if $\xi\|M(z) - M(z')\|^2 \leq \langle M(z) - M(z'), z - z'\rangle$, for every $(z, z') \in \mathcal{Z}^2$. The resolvent of a maximally monotone operator is 1-cocoercive and ∇G is $1/\nu$-cocoercive, for any ν-smooth function G.

The adjoint of a linear operator P is denoted by P^* and its operator norm by $\|P\|$. P is self-adjoint if $P = P^*$. Let $P : \mathcal{Z} \to \mathcal{Z}$ be a self-adjoint linear operator. P is positive if $\langle Pz, z\rangle \geq 0$, for every $z \in \mathcal{Z}$, and strongly positive if, additionally, $\langle Pz, z\rangle = 0$ implies $z = 0$. In this latter case, the inner product induced by P is defined by $\langle z, z'\rangle_P := \langle Pz, z'\rangle$ and the norm induced by P by $\|z\|_P := \langle z, z\rangle_P^{1/2}$. We denote by $\mathcal{Z}_P$ the real Hilbert space made from the vector space $\mathcal{Z}$ endowed with $\langle \cdot, \cdot\rangle_P$.

2 Primal–Dual Formulation and Optimality Conditions

For Problem (1) to be well posed, we suppose that there exists $x^\star \in \mathcal{X}$, such that

$$0 \in \nabla F(x^\star) + \partial R(x^\star) + L^* \partial H(Lx^\star). \tag{2}$$

Then, $x^\star$ is a solution to (1). For instance, a standard qualification constraint for this assumption to hold is that 0 belongs to the relative interior of $\mathrm{dom}(H) - L\,\mathrm{dom}(R)$ [20]. Then, for every $x^\star$ satisfying (2), there exists $y^\star \in \partial H(Lx^\star)$ such that $0 \in \nabla F(x^\star) + \partial R(x^\star) + L^* y^\star$; equivalently, $(x^\star, y^\star)$ is a zero of the set-valued operator

M defined by

$$M : (x, y) \in \mathcal{X} \times \mathcal{Y} \mapsto \Big(\nabla F(x) + \partial R(x) + L^* y, -Lx + \partial H^*(y)\Big). \tag{3}$$

Conversely, for every $(x^\star, y^\star) \in \operatorname{zer}(M)$, $x^\star$ is a solution to (1) and y^* is a solution to the dual problem

$$\underset{y \in \mathcal{Y}}{\text{minimize}} \Big((F + R)^*(-L^* y) + H^*(y)\Big), \tag{4}$$

see Sect. 15.3 of [5]; moreover, there exist $r^\star \in \partial R(x^\star)$ and $h^\star \in \partial H^*(y^\star)$ such that, using 2-block vector notations in $\mathcal{X} \times \mathcal{Y}$,

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \nabla F(x^\star) + r^\star + L^* y^\star \\ -Lx^\star + h^\star \end{bmatrix}. \tag{5}$$

In the sequel, we let $(x^\star, y^\star) \in \operatorname{zer}(M)$ and $r^\star, h^\star$ be any elements such that Eq. (5) holds.

A zero of M is also a saddle point of the convex–concave Lagrangian function, defined as

$$\mathscr{L}(x, y) := F(x) + R(x) - H^*(y) + \langle Lx, y \rangle. \tag{6}$$

For every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we define the *Lagrangian gap* at (x, y) as $\mathscr{L}(x, y^\star) - \mathscr{L}(x^\star, y)$. The following holds:

Lemma 2.1 *(Lagrangian gap) For every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have*

$$\mathscr{L}(x, y^\star) - \mathscr{L}(x^\star, y) = D_F(x, x^\star) + D_R(x, x^\star) + D_{H^*}(y, y^\star), \tag{7}$$

where the Bregman divergence of the smooth function F between any two points x and x' is $D_F(x, x') := F(x) - F(x') - \langle \nabla F(x'), x - x' \rangle$, and $D_R(x, x^\star) := R(x) - R(x^\star) - \langle r^\star, x - x^\star \rangle$, $D_{H^*}(y, y^\star) := H^*(y) - H^*(y^\star) - \langle h^\star, y - y^\star \rangle$.

Proof Using the optimality conditions (5), we have

$$\begin{aligned} D_F(x, x^\star) + D_R(x, x^\star) &= (F + R)(x) - (F + R)(x^\star) - \langle \nabla F(x^\star) + r^\star, x - x^\star \rangle \\ &= (F + R)(x) - (F + R)(x^\star) + \langle L^* y^\star, x - x^\star \rangle \\ &= (F + R)(x) - (F + R)(x^\star) + \langle y^\star, Lx \rangle - \langle y^\star, Lx^\star \rangle. \end{aligned}$$

We also have

$$\begin{aligned} D_{H^*}(y, y^\star) &= H^*(y) - H^*(y^\star) - \langle Lx^\star, y - y^\star \rangle \\ &= H^*(y) - H^*(y^\star) - \langle Lx^\star, y \rangle + \langle y^\star, Lx^\star \rangle. \end{aligned}$$

Hence,

$$
\begin{aligned}
&D_F(x,x^\star)+D_R(x,x^\star)+D_{H^*}(y,y^\star)\\
&\quad=(F+R)(x)-(F+R)(x^\star)+H^*(y)-H^*(y^\star)-\langle Lx^\star,y\rangle+\langle y^\star,Lx\rangle\\
&\quad=\mathscr{L}(x,y^\star)-\mathscr{L}(x^\star,y). \qquad\square
\end{aligned}
$$

For every $x\in\mathcal{X}$, $y\in\mathcal{Y}$, Lemma 2.1 and the convexity of F, R, H^* imply that $\mathscr{L}(x^\star,y)\le\mathscr{L}(x^\star,y^\star)\le\mathscr{L}(x,y^\star)$. So, the Lagrangian gap $\mathscr{L}(x,y^\star)-\mathscr{L}(x^\star,y)$ is nonnegative, and it is zero if x is a solution to Problem (1) and y is a solution to the dual problem (4). The converse is not always true, generally speaking. But in realistic situations, this is the case, and under mild assumptions, like strict convexity of the functions around $x^\star$ and $y^\star$, the Lagrangian gap converging to zero is a valid measure of convergence to a solution.

The operator M defined in (3) can be shown to be maximally monotone. Moreover, we have

$$
\begin{aligned}
M(x,y)&=\begin{bmatrix}\partial R(x)\\ 0\end{bmatrix}+\begin{bmatrix}L^*y\\ -Lx+\partial H^*(y)\end{bmatrix}+\begin{bmatrix}\nabla F(x)\\ 0\end{bmatrix} &(8)\\
&=\begin{bmatrix}0\\ \partial H^*(y)\end{bmatrix}+\begin{bmatrix}\partial R(x)+L^*y\\ -Lx\end{bmatrix}+\begin{bmatrix}\nabla F(x)\\ 0\end{bmatrix}, &(9)
\end{aligned}
$$

and each term at the right-hand side of (8) or (9) is maximally monotone, see Corollary 25.5 in [5].

Algorithm 1 *

Davis–Yin Splitting alg.
DYS($\tilde{A}, \tilde{B}, \tilde{C}$) [28]

```
Input: $v^0 \in \mathcal{Z}, \gamma > 0$
for $k = 0, 1, 2, \ldots$ do
    $z^k = J_{\gamma\tilde{B}}(v^k)$
    $u^{k+1} = J_{\gamma\tilde{A}}(2z^k - v^k - \gamma\tilde{C}(z^k))$
    $v^{k+1} = v^k + u^{k+1} - z^k$
end for
```

Algorithm 2 *

LiCoSGD (new)

```
Input: $x^0 \in \mathcal{X}, y^0 \in \mathcal{Y}, \gamma > 0, \tau > 0$
for $k = 0, 1, 2, \ldots$ do
    $w^k = x^k - \gamma g^{k+1}$
    $y^{k+1} = y^k + \tau L(w^k - \gamma L^* y^k) - \tau b$
    $x^{k+1} = w^k - \gamma L^* y^{k+1}$
end for
```

Note : the deterministic versions of the algorithms are obtained by setting $g^{k+1}=\nabla F(x^k)$.

Algorithm 3 *

```
Stochastic PDDY alg. (new)
Input: $p^0 \in \mathcal{X}, y^0 \in \mathcal{Y}, \gamma > 0, \tau > 0$
for $k = 0, 1, 2, \ldots$ do
    $y^{k+1} = \mathrm{prox}_{\tau H^*}\big(y^k + \tau L(p^k - \gamma L^* y^k)\big)$
    $x^k = p^k - \gamma L^* y^{k+1}$
    $s^{k+1} = \mathrm{prox}_{\gamma R}\left(2x^k - p^k - \gamma g^{k+1}\right)$
    $p^{k+1} = p^k + s^{k+1} - x^k$
end for
```

Algorithm 4 *

```
Stochastic PD3O alg. (new)
Input: $p^0 \in \mathcal{X}, y^0 \in \mathcal{Y}, \gamma > 0, \tau > 0$
for $k = 0, 1, 2, \ldots$ do
    $x^k = \mathrm{prox}_{\gamma R}(p^k)$
    $w^k = 2x^k - p^k - \gamma g^{k+1}$
    $y^{k+1} = \mathrm{prox}_{\tau H^*}\big(y^k + \tau L(w^k - \gamma L^* y^k)\big)$
    $p^{k+1} = x^k - \gamma g^{k+1} - \gamma L^* y^{k+1}$
end for
```

3 Davis–Yin Splitting

Solving Problem (1) boils down to finding a zero $(x^\star, y^\star)$ of the monotone operator M defined in (3), which can be written as the sum of three monotone operators, like in (8) or (9). The method proposed by Davis and Yin [28], which we call Davis–Yin splitting (DYS), is dedicated to this problem; that is, find a zero of the sum of three monotone operators, one of which is cocoercive.

Let $\mathcal{Z}$ be a real Hilbert space. Let $\tilde{A}, \tilde{B}, \tilde{C}$ be maximally monotone operators on $\mathcal{Z}$. We assume that $\tilde{C}$ is ξ-cocoercive, for some $\xi > 0$. The DYS algorithm, denoted by DYS$(\tilde{A}, \tilde{B}, \tilde{C})$ and shown above, aims at finding an element in $\mathrm{zer}(\tilde{A} + \tilde{B} + \tilde{C})$, supposed nonempty. The fixed points of DYS$(\tilde{A}, \tilde{B}, \tilde{C})$ are the triplets $(v^\star, z^\star, u^\star) \in \mathcal{Z}^3$, such that

$$z^\star = J_{\gamma \tilde{B}}(v^\star), \quad u^\star = J_{\gamma \tilde{A}}\big(2z^\star - v^\star - \gamma \tilde{C}(z^\star)\big), \quad u^\star = z^\star. \tag{10}$$

These fixed points are related to the zeros of $\tilde{A} + \tilde{B} + \tilde{C}$ as follows, see Lemma 2.2 in [28]: for every $(v^\star, z^\star, u^\star) \in \mathcal{Z}^3$ satisfying (10), $z^\star \in \mathrm{zer}(\tilde{A} + \tilde{B} + \tilde{C})$. Conversely, for every $z^\star \in \mathrm{zer}(\tilde{A} + \tilde{B} + \tilde{C})$, there exists $(v^\star, u^\star) \in \mathcal{Z}^2$, such that $(v^\star, z^\star, u^\star)$ satisfies (10). We have [28]:

Lemma 3.1 *[Convergence of the DYS algorithm] Suppose that $\gamma \in (0, 2\xi)$. Then, the sequences $(v^k)_{k\in\mathbb{N}}$, $(z^k)_{k\in\mathbb{N}}$, $(u^k)_{k\in\mathbb{N}}$ generated by DYS$(\tilde{A}, \tilde{B}, \tilde{C})$ converge to some elements $v^\star$, $z^\star$, $u^\star$ in $\mathcal{Z}$, respectively. Moreover, $(v^\star, z^\star, u^\star)$ satisfies (10) and $u^\star = z^\star \in \mathrm{zer}(\tilde{A} + \tilde{B} + \tilde{C})$.*

The following equality is at the heart of the convergence proofs:

Lemma 3.2 *[Fundamental equality of the DYS algorithm] Let $(v^k, z^k, u^k) \in \mathcal{Z}^3$ be the iterates of the DYS algorithm, and $(v^\star, z^\star, u^\star) \in \mathcal{Z}^3$ be such that (10) holds. Then, for every $k \geq 0$, there exist $b^k \in \tilde{B}(z^k)$, $b^\star \in \tilde{B}(z^\star)$, $a^{k+1} \in \tilde{A}(u^{k+1})$ and $a^\star \in \tilde{A}(u^\star)$, such that*

$$\begin{aligned}\|v^{k+1} - v^\star\|^2 = \|v^k - v^\star\|^2 &- 2\gamma\langle b^k - b^\star, z^k - z^\star\rangle - 2\gamma\langle a^{k+1} - a^\star, u^{k+1} - u^\star\rangle \\ &- 2\gamma\langle \tilde{C}(z^k) - \tilde{C}(z^\star), z^k - z^\star\rangle + \gamma^2\|\tilde{C}(z^k) - \tilde{C}(z^\star)\|^2 \\ &- \gamma^2\|a^{k+1} + b^k - a^\star - b^\star\|^2.\end{aligned} \tag{11}$$

Proof Since $z^k = J_{\gamma\tilde{B}}(v^k)$, $z^k \in v^k - \gamma\tilde{B}(z^k)$ by definition of the resolvent. Therefore, there exists $b^k \in \tilde{B}(z^k)$, such that $z^k = v^k - \gamma b^k$. Similarly, $u^{k+1} \in 2z^k - v^k - \gamma\tilde{C}(z^k) - \gamma\tilde{A}(u^{k+1}) = v^k - 2\gamma b^k - \gamma\tilde{C}(z^k) - \gamma\tilde{A}(u^{k+1})$. Therefore, there exists $a^{k+1} \in \tilde{A}(u^{k+1})$, such that

$$\begin{cases} z^k = v^k - \gamma b^k \\ u^{k+1} = v^k - 2\gamma b^k - \gamma\tilde{C}(z^k) - \gamma a^{k+1} \\ v^{k+1} = v^k + u^{k+1} - z^k. \end{cases} \tag{12}$$

Moreover, $v^{k+1} = v^k - \gamma b^k - \gamma\tilde{C}(z^k) - \gamma a^{k+1}$. Similarly, there exist $a^\star \in \tilde{A}(u^\star)$ and $b^\star \in \tilde{B}(z^\star)$, such that

$$\begin{cases} z^\star = v^\star - \gamma b^\star \\ u^\star = v^\star - 2\gamma b^\star - \gamma\tilde{C}(z^\star) - \gamma a^\star \\ v^\star = v^\star + u^\star - z^\star, \end{cases} \tag{13}$$

and
$v^\star = v^\star - \gamma b^\star - \gamma\tilde{C}(z^\star) - \gamma a^\star$. Therefore,

$$\begin{aligned} \|v^{k+1} - v^\star\|^2 = \|v^k - v^\star\|^2 - 2\gamma\Big\langle a^{k+1} + b^k + \tilde{C}(z^k) - \big(a^\star + b^\star + \tilde{C}(z^\star)\big), \\ v^k - v^\star\Big\rangle + \gamma^2\|a^{k+1} + b^k + \tilde{C}(z^k) - \big(a^\star + b^\star + \tilde{C}(z^\star)\big)\|^2. \end{aligned}$$

By expanding the last squared norm and by using (12) and (13) in the inner product, we get

$$\begin{aligned} \|v^{k+1} - v^\star\|^2 &= \|v^k - v^\star\|^2 - 2\gamma\langle b^k + \tilde{C}(z^k) - \big(b^\star + \tilde{C}(z^\star)\big), z^k - z^\star\rangle \\ &\quad - 2\gamma\langle a^{k+1} - a^\star, u^{k+1} - u^\star\rangle \\ &\quad - 2\gamma\langle b^k + \tilde{C}(z^k) - \big(b^\star + \tilde{C}(z^\star)\big), \gamma b^k - \gamma b^\star\rangle \\ &\quad - 2\gamma\langle a^{k+1} - a^\star, 2\gamma b^k + \gamma\tilde{C}(z^k) + \gamma a^{k+1} \\ &\quad - \big(2\gamma b^\star + \gamma\tilde{C}(z^\star) + \gamma a^\star\big)\rangle \\ &\quad + \gamma^2\|a^{k+1} + b^k - \big(a^\star + b^\star\big)\|^2 + \gamma^2\|\tilde{C}(z^k) - \tilde{C}(z^\star)\|^2 \\ &\quad + 2\gamma^2\langle a^{k+1} + b^k - \big(a^\star + b^\star\big), \tilde{C}(z^k) - \tilde{C}(z^\star)\rangle \\ &= \|v^k - v^\star\|^2 - 2\gamma\langle b^k - b^\star, z^k - z^\star\rangle - 2\gamma\langle a^{k+1} - a^\star, u^{k+1} - u^\star\rangle \\ &\quad - 2\gamma\langle\tilde{C}(z^k) - \tilde{C}(z^\star), z^k - z^\star\rangle + \gamma^2\|\tilde{C}(z^k) - \tilde{C}(z^\star)\|^2 \\ &\quad - 2\gamma^2\|b^k - b^\star\|^2 - 2\gamma^2\langle a^{k+1} - a^\star, 2b^k + a^{k+1} - 2b^\star - a^\star\rangle \\ &\quad + \gamma^2\|a^{k+1} + b^k - \big(a^\star + b^\star\big)\|^2. \end{aligned}$$

After combining the last three terms into $-\gamma^2\|a^{k+1} + b^k - \big(a^\star + b^\star\big)\|^2$, we obtain the result. □

4 A Class of Four Primal–Dual Optimization Algorithms

We now set $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the spaces defined in Sect. 2. To solve the primal–dual problem (8) or (9), which consists in finding a zero of the sum $A + B + C$ of three operators in $\mathcal{Z}$, of which C is cocoercive, a natural idea is to apply the Davis–Yin algorithm $\mathrm{DYS}(A, B, C)$. But the resolvent of A or B is often intractable. In this section, we show that preconditioning is the solution; that is, we exhibit a strongly positive linear operator P, such that $\mathrm{DYS}(P^{-1}A, P^{-1}B, P^{-1}C)$ is tractable. Since $P^{-1}A$, $P^{-1}B$, $P^{-1}C$ are monotone operators in $\mathcal{Z}_P$, the algorithm will converge to a zero of $P^{-1}A + P^{-1}B + P^{-1}C$, or, equivalently, of $A + B + C$. Let us apply this idea in four different ways.

4.1 A New Primal–Dual Algorithm: The PDDY Algorithm

Let $\gamma > 0$ and $\tau > 0$ be real parameters. We introduce the four operators on $\mathcal{Z}$, using matrix-vector notations:

$$A(x, y) = \begin{bmatrix} L^* y \\ -Lx + \partial H^*(y) \end{bmatrix}, \quad B(x, y) = \begin{bmatrix} \partial R(x) \\ 0 \end{bmatrix}, \; C(x, y) = \begin{bmatrix} \nabla F(x) \\ 0 \end{bmatrix},$$
$$P = \begin{bmatrix} I & 0 \\ 0 & \frac{\gamma}{\tau} I - \gamma^2 LL^* \end{bmatrix}. \tag{14}$$

P is strongly positive if and only if $\gamma\tau\|L\|^2 < 1$. Since A, B, C are maximally monotone in $\mathcal{Z}$, $P^{-1}A$, $P^{-1}B$, $P^{-1}C$ are maximally monotone in $\mathcal{Z}_P$. Moreover, $P^{-1}C$ is $1/\nu$-cocoercive in $\mathcal{Z}_P$. Importantly, we have:

$$P^{-1}C : (x, y) \mapsto \big(\nabla F(x), 0\big), \qquad J_{\gamma P^{-1}B} : (x, y) \mapsto \big(\mathrm{prox}_{\gamma R}(x), y\big),$$
$$J_{\gamma P^{-1}A} : (x, y) \mapsto (x', y'), \qquad \text{where} \left\lfloor \begin{array}{l} y' = \mathrm{prox}_{\tau H^*}\big(y + \tau L(x - \gamma L^* y)\big) \\ x' = x - \gamma L^* y'. \end{array} \right. \tag{15}$$

The form of the last resolvent was shown in [55]; see also [25], where this resolvent appears as one iteration of the Proximal Method of Multipliers. We plug these explicit steps into the Davis–Yin algorithm $\mathrm{DYS}(P^{-1}B, P^{-1}A, P^{-1}C)$ and we identify the variables as $v^k = (p^k, q^k)$, $z^k = (x^k, y^{k+1})$, $u^k = (s^k, d^k)$, for some variables $(p^k, x^k, s^k) \in \mathcal{X}^3$ and $(q^k, y^k, d^k) \in \mathcal{Y}^3$. Thus, we do the following substitutions:

- Using (15), the step $z^k = J_{\gamma P^{-1}A}(v^k)$, is equivalent to

$$\left\lfloor \begin{array}{l} y^{k+1} = \mathrm{prox}_{\tau H^*}\big((I - \tau\gamma LL^*)q^k + \tau L p^k\big) \\ x^k = p^k - \gamma L^* y^{k+1}. \end{array} \right.$$

- The step $u^{k+1} = J_{\gamma P^{-1}B}\big(2z^k - v^k - \gamma P^{-1}C(z^k)\big)$ is equivalent to

$$\left\lfloor \begin{array}{l} s^{k+1} = \mathrm{prox}_{\gamma R}\big(2x^k - p^k - \gamma \nabla F(x^k)\big) \\ d^{k+1} = 2y^{k+1} - q^k. \end{array} \right.$$

- Finally, the step $v^{k+1} = v^k + u^{k+1} - z^k$ is equivalent to

$$\left\lfloor \begin{array}{l} p^{k+1} = p^k + s^{k+1} - x^k \\ q^{k+1} = q^k + d^{k+1} - y^{k+1}. \end{array} \right.$$

We can replace q^k by y^k and discard d^k, which is not needed. This yields the new Primal–Dual Davis–Yin (PDDY) algorithm, shown above (with $g^{k+1} = \nabla F(x^k)$). Note that it can be written with only one call to L and L^* per iteration. Also, the PDDY Algorithm could be overrelaxed [25], since this possibility exists for the Davis–Yin algorithm. We have:

Theorem 4.1 *[Convergence of the PDDY Algorithm] Suppose that $\gamma \in (0, 2/\nu)$ and that $\tau\gamma\|L\|^2 < 1$. Then, the sequences $(x^k)_{k\in\mathbb{N}}$ and $(s^k)_{k\in\mathbb{N}}$ generated by the PDDY Algorithm converge to the same solution $x^\star$ to Problem (1), and the sequence $(y^k)_{k\in\mathbb{N}}$ converges to some dual solution $y^\star$ of (4).*

Proof Under the assumptions of Theorem 4.1, P is strongly positive. Then, the result follows from Lemma 3.1 applied in $\mathcal{Z}_P$ and from the analysis in Sect. 2. □

4.2 The PD3O Algorithm

We consider the same notations as in the previous section. We switch the roles of A and B and consider DYS$(P^{-1}A, P^{-1}B, P^{-1}C)$. Then, after some substitutions similar to the ones done to construct the PDDY algorithm, we recover exactly the PD3O algorithm proposed in [77]. Although it is not derived this way, its interpretation as a primal–dual Davis–Yin algorithm is mentioned by its author. Its convergence properties are the same as for the PDDY Algorithm, as stated in Theorem 4.1.

In a recent work [55], the PD3O algorithm has been shown to be an instance of the Davis–Yin algorithm, with a different reformulation, which does not involve duality. The authors of the present paper developed this technique further, applied it to the PDDY algorithm as well and obtained convergence rates and accelerations for both algorithms [26].

4.3 The Condat–Vũ Algorithm

Let $\gamma > 0$ and $\tau > 0$ be real parameters. We want to study the decomposition (9) instead of (8). For this, we define the operators

$$\bar{A}(x, y) = \begin{bmatrix} \partial R(x) \ + L^* y \\ -Lx \end{bmatrix}, \ \bar{B}(x, y) = \begin{bmatrix} 0 \\ \partial H^*(y) \end{bmatrix}, \ Q = \begin{bmatrix} K & 0 \\ 0 & I \end{bmatrix}, \tag{16}$$

where $K := \frac{\gamma}{\tau} I - \gamma^2 L^* L$, and we define C like in (14). If $\gamma\tau\|L\|^2 < 1$, K and Q are strongly positive. In that case, since $\bar{A}, \bar{B}, C$ are maximally monotone in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Q^{-1}\bar{A}$, $Q^{-1}\bar{B}$, $Q^{-1}C$ are maximally monotone in $\mathcal{Z}_Q$. Moreover, we have:

$$Q^{-1}C:(x,y)\mapsto\big(K^{-1}\nabla F(x),0\big),\quad J_{\gamma Q^{-1}\bar{B}}:(x,y)\mapsto\big(x,\operatorname{prox}_{\gamma H^*}(y)\big),\qquad(17)$$
$$J_{\gamma Q^{-1}\bar{A}}:(x,y)\mapsto(x',y'),\text{ where }\left\lfloor\begin{array}{l}x'=\operatorname{prox}_{\tau R}\big((I-\tau\gamma L^*L)x-\tau L^*y\big)\\ y'=y+\gamma Lx'.\end{array}\right.$$

If we plug these explicit steps into the Davis–Yin algorithm DYS$(Q^{-1}\bar{A}, Q^{-1}\bar{B}, Q^{-1}C)$ or DYS$(Q^{-1}\bar{B}, Q^{-1}\bar{A}, Q^{-1}C)$, and after straightforward simplifications, we recover the two forms of the Condat–Vũ algorithm [22, 72]; that is, Algorithms 3.1 and 3.2 of [22], respectively, see also in [25]. The Condat–Vũ algorithm has the form of a primal–dual forward–backward algorithm [16, 25, 44]. We have just shown that it can be viewed as a primal–dual Davis–Yin algorithm, with a different metric, as well. Furthermore, it is easy to show that $Q^{-1}C$ is ξ-cocoercive, with $\xi=(\frac{\gamma}{\tau}-\gamma^2\|L\|^2)/\nu$. Hence, convergence follows from Lemma 3.1, under the same condition on τ and γ as in Theorem 3.1 of [22], namely $\frac{\nu}{2}<\frac{1}{\tau}-\gamma\|L\|^2$.

5 Stochastic Primal–Dual Algorithms

We now introduce stochastic versions of the PD3O and PDDY algorithms; we omit the analysis of stochastic versions of the Condat–Vũ algorithm, which is the same, with added technicalities due to cocoercivity with respect to the metric induced by Q in (16). Our approach has a ‘plug-and-play’ flavor: we show that we have all the ingredients to leverage the unified theory of stochastic gradient estimators recently presented in [36].

In the stochastic versions of the algorithms, the gradient $\nabla F(x^k)$ is replaced by a stochastic gradient g^{k+1}. That is, we consider a filtered probability space $(\Omega,\mathscr{F},(\mathscr{F}_k)_{k\in\mathbb{N}},\mathbb{P})$, an $(\mathscr{F}_k)$-adapted stochastic process $(g^k)_{k\in\mathbb{N}}$, we denote by $\mathbb{E}$ the expectation and by $\mathbb{E}_k$ the conditional expectation w.r.t. $\mathscr{F}_k$. The following assumption is made on the process $(g^k)_{k\in\mathbb{N}}$.

Assumption 1 There exist $\alpha,\beta,\delta\geq 0$, $\rho\in(0,1]$ and a $(\mathscr{F}_k)_{k\in\mathbb{N}}$-adapted stochastic process denoted by $(\sigma_k)_{k\in\mathbb{N}}$, such that, for every $k\in\mathbb{N}$, $\mathbb{E}_k(g^{k+1})=\nabla F(x^k)$, $\mathbb{E}_k(\|g^{k+1}-\nabla F(x^\star)\|^2)\leq 2\alpha D_F(x^k,x^\star)+\beta\sigma_k^2$, and $\mathbb{E}_k(\sigma_{k+1}^2)\leq(1-\rho)\sigma_k^2+2\delta D_F(x^k,x^\star)$.

Assumption 1 is satisfied by several stochastic gradient estimators used in machine learning, including some types of coordinate descent [73], variance reduction [38], and also compressed gradients used to reduce the communication cost in distributed optimization [4, 65, 75], see Table 1 in [36]. Also, the full gradient estimator defined by $g^{k+1}=\nabla F(x^k)$ satisfies Assumption 1 with $\alpha=\nu$, the smoothness constant of F, $\sigma_k\equiv 0$, $\rho=1$, and $\delta=\beta=0$, see Theorem 2.1.5 in [54]. The loopless SVRG estimator [39, 45] also satisfies Assumption 1:

Proposition 5.1 *(Loopless SVRG estimator) Assume that F is written as a sum $F=\frac{1}{n}\sum_{i=1}^n f_i$, for some $n\geq 1$, where for every $i\in\{1,\ldots,n\}$, $f_i:\mathcal{X}\to\mathbb{R}$ is a ν_i-smooth convex function. Let $p\in(0,1)$, and $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space. On $(\Omega,\mathscr{F},\mathbb{P})$, consider:*

- *a sequence of i.i.d. random variables* $(\theta^k)_{k\in\mathbb{N}}$ *with Bernoulli distribution of parameter* p*,*
- *a sequence of i.i.d random variables* $(\zeta^k)_{k\in\mathbb{N}}$ *with uniform distribution over* $\{1,\dots,n\}$*,*
- *the sigma-field* $\mathscr{F}_k$ *generated by* $(\theta^k,\zeta^k)_{0\le j\le k}$ *and a* $(\mathscr{F}_k)$*-adapted stochastic process* $(x^k)_{k\in\mathbb{N}}$*,*
- *a stochastic process* $(\tilde{x}^k)_{k\in\mathbb{N}}$ *defined by* $\tilde{x}^{k+1}=\theta^{k+1}x^k+(1-\theta^{k+1})\tilde{x}^k$*,*
- *a stochastic process* $(g^k)_{k\in\mathbb{N}}$ *defined by* $g^{k+1}=\nabla f_{\zeta^{k+1}}(x^k)-\nabla f_{\zeta^{k+1}}(\tilde{x}^k)+\nabla F(\tilde{x}^k)$*.*

Then, the process $(g^k)_{k\in\mathbb{N}}$ *satisfies Assumption 1 with* $\alpha=2\max_{i\in\{1,\dots,n\}}\nu_i$*,* $\beta=2$*,* $\rho=p$*,* $\delta=\alpha p/2$*, and* $\sigma_k^2=\frac{1}{n}\sum_{i=1}^n\mathbb{E}_k\|\nabla f_i(\tilde{x}^k)-\nabla f_i(x^\star)\|^2$*.*

Proof The proof is the same as the proof of Lemma A.11 of [36], which is only stated for $(x^k)_{k\in\mathbb{N}}$ generated by a specific algorithm, but remains true for any $(\mathscr{F}_k)$-adapted stochastic process $(x^k)_{k\in\mathbb{N}}$. □

We can now exhibit our main results. In a nutshell, $P^{-1}C(z^k)$ is replaced by the random realization $P^{-1}(g^{k+1},0)$ and the last term of Eq. (11), which is nonnegative, is handled using Assumption 1.

5.1 The Stochastic PD3O Algorithm

The Stochastic PD3O Algorithm, shown above, has $\mathcal{O}(1/k)$ ergodic convergence in the general case. A linear convergence result in the strongly convex setting is derived in Appendix 1.

Theorem 5.1 *[Convergence of the Stochastic PD3O Algorithm] Suppose that Assumption 1 holds. Let* $\kappa:=\beta/\rho$*,* $\gamma,\tau>0$ *be such that* $\gamma\le 1/2(\alpha+\kappa\delta)$ *and* $\gamma\tau\|L\|^2<1$*. Set* $V^0:=\|v^0-v^\star\|_P^2+\gamma^2\kappa\sigma_0^2$*, where* $v^0=(p^0,y^0)$*. Then, for every* $k\in\mathbb{N}$*,*

$$\mathbb{E}\left(\mathscr{L}(\bar{x}^k,y^\star)-\mathscr{L}(x^\star,\bar{y}^{k+1})\right)\le\frac{V^0}{k\gamma},$$

where $\bar{x}^k=\frac{1}{k}\sum_{j=0}^{k-1}x^j$ *and* $\bar{y}^{k+1}=\frac{1}{k}\sum_{j=1}^{k}y^j$*.*

Proof Using Lemma A.1, the convexity of F, R, H^*, and Lemma 2.1,

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2+\kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2&\le\|v^k-v^\star\|_P^2+\kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\&\quad-2\gamma(1-\gamma(\alpha+\kappa\delta))\mathbb{E}_k\left(\mathscr{L}(x^k,d^\star)-\mathscr{L}(x^\star,d^{k+1})\right).\end{aligned}$$

We have $1-\rho+\beta/\kappa=1$, $\gamma\le 1/2(\alpha+\kappa\delta)$. Set $V^k:=\|v^k-v^\star\|_P^2+\kappa\gamma^2\sigma_k^2$, for every $k\in\mathbb{N}$. Then, $\mathbb{E}_kV^{k+1}\le V^k-\gamma\mathbb{E}_k\left(\mathscr{L}(x^k,d^\star)-\mathscr{L}(x^\star,d^{k+1})\right)$. Taking the expectation, $\gamma\mathbb{E}\left(\mathscr{L}(x^k,d^\star)-\mathscr{L}(x^\star,d^{k+1})\right)\le\mathbb{E}V^k-\mathbb{E}V^{k+1}$. Iterating and using the nonnegativity of V^k, $\gamma\sum_{j=0}^{k-1}\mathbb{E}\left(\mathscr{L}(x^j,d^\star)-\mathscr{L}(x^\star,d^{j+1})\right)\le\mathbb{E}V^0$. Finally, note that $d^{k+1} = y^{k+1}$ and $d^\star = y^\star$. Indeed, $y^k = q^k$ and $q^{k+1} = q^k + d^{k+1} - y^k = d^{k+1}$. We can conclude using the convex-concavity of $\mathscr{L}$. □

In the deterministic case $g^{k+1} = \nabla F(x^k)$, we recover the same rate as in [77][Theorem 2].

Remark 5.1 (Primal–Dual gap) Deriving a similar bound on the stronger primal–dual gap $(F+R+H\circ L)(\bar{x}^k)+\big((F+R)^*\circ(-L)+H^*\big)(\bar{y}^k)$ requires additional assumptions; for instance, even for the Chambolle–Pock algorithm, which is the particular case of the PD3O, PPDY and Condat–Vũ algorithms when $F = 0$, the best available result [13][Theorem 1] is not stronger than Theorem 5.1

Remark 5.2 (Particular case of SGD) In the case where $H = 0$ and $L = 0$, the Stochastic PD3O Algorithm boils down to proximal stochastic gradient descent (SGD) and Theorem 5.1 implies that $\mathbb{E}\left((F+R)(\bar{x}^k) - (F+R)(x^\star)\right) \leq V^0/(\gamma k)$.

This $\mathcal{O}(1/k)$ ergodic convergence rate unifies known results on SGD in the non-strongly-convex case, whenever the stochastic gradient satisfies Assumption 1.

5.2 The Stochastic PDDY Algorithm

We now analyze the proposed Stochastic PDDY Algorithm, shown above. For it too, we have $\mathcal{O}(1/k)$ ergodic convergence in the general case. A linear convergence result in the strongly convex setting is derived in Appendix 1.

Theorem 5.2 *(Convergence of the Stochastic PDDY Algorithm) Suppose that Assumption 1 holds. Let $\kappa := \beta/\rho$, $\gamma, \tau > 0$ be such that $\gamma \leq 1/2(\alpha + \kappa\delta)$ and $\gamma\tau\|L\|^2 < 1$. Define $V^0 := \|v^0 - v^\star\|_P^2 + \gamma^2\kappa\sigma_0^2$, where $v^0 = (p^0, y^0)$. Then, for every $k \in \mathbb{N}$,*

$$\mathbb{E}\left(D_F(\bar{x}^k, x^\star) + D_{H^*}(\bar{y}^{k+1}, y^\star) + D_R(\bar{s}^{k+1}, s^\star)\right) \leq \frac{V^0}{k\gamma},$$

where $\bar{x}^k = \frac{1}{k}\sum_{j=0}^{k-1} x^j$, $\bar{y}^{k+1} = \frac{1}{k}\sum_{j=1}^{k} y^j$ and $\bar{s}^{k+1} = \frac{1}{k}\sum_{j=1}^{k} s^j$.

Proof Using Lemma A.2 and the convexity of F, R, H^*,

$$\mathbb{E}_k\|v^{k+1} - v^\star\|_P^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 \leq \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2$$
$$- 2\gamma\big(1 - \gamma(\alpha + \kappa\delta)\big)\left(D_F(x^k, x^\star) + D_{H^*}(y^k, y^\star) + \mathbb{E}_k D_R(s^{k+1}, s^\star)\right).$$

Since $1 - \rho + \beta/\kappa = 1$, $\gamma \leq 1/2(\alpha + \kappa\delta)$. Set $V^k := \|v^k - v^\star\|_P^2 + \kappa\gamma^2\sigma_k^2$. Then,

$$\mathbb{E}_k V^{k+1} \leq V^k - \gamma\mathbb{E}_k\left(D_F(x^k, x^\star) + D_{H^*}(y^k, y^\star) + D_R(s^{k+1}, s^\star)\right).$$

Taking the expectation, $\gamma\mathbb{E}\left(D_F(x^k, x^\star) + D_{H^*}(y^k, y^\star) + D_R(s^{k+1}, s^\star)\right) \leq \mathbb{E}V^k - \mathbb{E}V^{k+1}$. Iterating and using the nonnegativity of V^k, $\gamma\sum_{j=0}^{k-1}\mathbb{E}\big(D_F(x^k, x^\star) + D_{H^*}(y^k, y^\star) + D_R(s^{k+1}, s^\star)\big) \leq \mathbb{E}V^0$. We conclude using the convexity of the Bregman divergence in its first variable. □

6 Linearly Constrained Smooth Optimization

In this section, we consider the problem

$$\operatorname*{minimize}_{x\in\mathcal{X}} \; F(x) \quad \text{s.t.} \quad Lx = b, \tag{18}$$

where $L : \mathcal{X} \to \mathcal{Y}$ is a linear operator, $\mathcal{X}$ and $\mathcal{Y}$ are real Hilbert spaces, F is a ν-smooth convex function, for some $\nu > 0$, and $b \in \operatorname{ran}(L)$, the range of L. This is a particular case of Problem (1) with $R = 0$ and

$H = \iota_b$. We suppose that a solution $x^\star$ exists, satisfying $Lx^\star = b$ and $0 \in \nabla F(x^\star) + L^* y^\star$ for some $y^\star \in \mathcal{Y}$. The stochastic PD3O and PDDY algorithms both revert to the same algorithm, shown above, which we call Linearly Constrained Stochastic Gradient Descent (LiCoSGD). It is fully split: it does not make use of projections onto the affine space $\{x \in \mathcal{X}, Lx = b\}$ and only makes calls to L and L^*.

In the deterministic case $g^{k+1} = \nabla F(x^k)$, LiCoSGD reverts to an instance of the algorithm first proposed by Loris and Verhoeven in [52] and rediscovered independently as the PDFP2O algorithm [15] and the Proximal Alternating Predictor–Corrector (PAPC) algorithm [30]. Convergence of this algorithm follows from Theorem 4.1, see other results in [25, 26]. Thus, LiCoSGD is a stochastic extension of this algorithm, for which Theorem 5.1 becomes:

Theorem 6.1 *(Convergence of LiCoSGD) Suppose that Assumption 1 holds. Let* $\kappa := \beta/\rho$, $\gamma, \tau > 0$ *be such that* $\gamma \le 1/2(\alpha + \kappa\delta)$ *and* $\gamma\tau\|L\|^2 < 1$. *Set* $V^0 := \|v^0 - v^\star\|_P^2 + \gamma^2\kappa\sigma_0^2$, *where* $v^0 = (w^0, y^0)$. *Then, for every* $k \in \mathbb{N}$,

$$\mathbb{E}\left(F(\bar{x}^k) - F(x^\star) + \langle L\bar{x}^k - b, y^\star\rangle\right) \le \frac{V^0}{k\gamma}, \tag{19}$$

where $\bar{x}^k = \frac{1}{k}\sum_{j=0}^{k-1} x^j$, $x^\star$ *and* $y^\star$ *are some primal and dual solutions.*

The convex function $x \mapsto F(x) - F(x^\star) + \langle Lx - b, y^\star\rangle$ is nonnegative and its minimum is zero, attained at $x^\star$. Under additional assumptions, like strict convexity around $x^\star$, this function takes value zero only if $F(x) = F(x^\star)$ and $Lx = b$, so that x is a solution.

We now state an important result: strong convexity of F is sufficient to get linear convergence. We denote by $\omega(W)$ the smallest positive eigenvalue of a positive self-adjoint linear operator W. Then, it is easy to show that for every $y \in \operatorname{ran}(L)$, $\omega(LL^*)\|y\|^2 \le \|L^* y\|^2$. Also, $\omega(LL^*) = \omega(L^*L)$.

Theorem 6.2 *(Linear convergence of LiCoSGD with F strongly convex) Suppose that Assumption 1 holds, that* F *is* μ_F*-strongly convex, for some* $\mu_F > 0$, *and that* $y^0 \in \operatorname{ran}(L)$. *Let* $x^\star$ *be the unique solution of (18),* $y^\star$ *be the unique element of* $\operatorname{ran}(L)$ *such that* $\nabla F(x^\star) + L^* y^\star = 0$. *Suppose that* $\gamma > 0$ *and* $\tau > 0$ *are such that* $\gamma\tau\|L\|^2 < 1$ *and* $\gamma \le 1/\alpha + \kappa\delta$, *for some* $\kappa > \beta/\rho$. *Define, for every* $k \in \mathbb{N}$,

$$V^k := \|x^k - x^\star\|^2 + \left(1 + \tau\gamma\omega(L^*L)\right)\|y^k - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\mathbb{E}\sigma_k^2, \tag{20}$$

and

$$r := \max\left(1-\gamma\mu_F, 1-\rho+\frac{\beta}{\kappa}, \frac{1}{1+\tau\gamma\omega(L^*L)}\right) < 1. \tag{21}$$

Then, for every $k \in \mathbb{N}$, $\mathbb{E}V^k \le r^k V^0$.

Proof Noting that $y^\star = d^\star = q^\star$ and applying Lemma A.1 with $\gamma \le (\alpha + \kappa\delta)$,

$$\mathbb{E}_k\|p^{k+1}-p^\star\|^2 + \mathbb{E}_k\|q^{k+1}-q^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}_k\sigma^2_{k+1} \le \|p^k-p^\star\|^2 + \|q^k-q^\star\|^2_{\gamma,\tau} - \gamma\mu_F\|x^k-x^\star\|^2 + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2 - \gamma^2\|P^{-1}A(u^{k+1}) - P^{-1}A(u^\star)\|^2_P.$$

Since the component of $P^{-1}A(u^{k+1}) - P^{-1}A(u^\star)$ in $\mathcal{X}$ is $L^*d^{k+1} - L^*d^\star$, we have

$$\mathbb{E}_k\|p^{k+1}-p^\star\|^2 + \mathbb{E}_k\|q^{k+1}-q^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}_k\sigma^2_{k+1} \le \|x^k-x^\star\|^2 + \|q^k-q^\star\|^2_{\gamma,\tau} - \gamma\mu_F\|p^k-p^\star\|^2 + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2 - \gamma^2\|L^*d^{k+1} - L^*d^\star\|^2.$$

Inspecting the iterations of the algorithm, one can see that $d^0 \in \operatorname{ran}(L)$ implies $d^{k+1} \in \operatorname{ran}(L)$. Since $d^\star \in \operatorname{ran}(L)$, $d^{k+1} - d^\star \in \operatorname{ran}(L)$. Therefore, $\omega(LL^*)\|d^{k+1}-d^\star\|^2 \le \|L^*d^{k+1} - L^*d^\star\|^2$. Since $q^{k+1} = d^{k+1} = y^{k+1}$ and $x^k = p^k$,

$$\mathbb{E}_k\|x^{k+1}-x^\star\|^2 + (1+\gamma\tau\omega(LL^*))\mathbb{E}_k\|y^{k+1}-y^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}_k\sigma^2_{k+1} \le (1-\gamma\mu_F)\|x^k-x^\star\|^2 + \|y^k-y^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2.$$

Thus, by setting V^k as in (20) and r as in (21), we have $\mathbb{E}_k V^{k+1} \le rV^k$. □

To the best of our knowledge, even in the deterministic case (with $\alpha = \nu$, $\rho = 1$, $\delta = \beta = 0$, $\kappa = 1$), this is a first time that a fully split algorithm using ∇F, L and L^* is shown to converge linearly to a solution of (18), whenever F is strongly convex. Also, the knowledge of μ_F is not needed. We discuss the application of LiCoSGD to decentralized optimization in Appendix 1.

7 Experiments

We present numerical experiments for the PDDY, PD3O and Condat–Vũ (CV) [22][Algorithm 3.1] algorithms. We observed that the performance of these algorithms is nearly identical, when the same stepsizes are used, but the PDDY and PD3O algorithms have a larger range of stepsizes than the CV algorithm, so that they are often faster after tuning. We used $\gamma\tau\|L\|^2 = 0.999$, which was always the best choice for these two algorithms. So, we do not provide direct comparisons in the plots. Instead, we focus on how the choice of the stochastic gradient estimator affects the convergence speed; we compare the true gradient, the standard stochastic gradient estimator

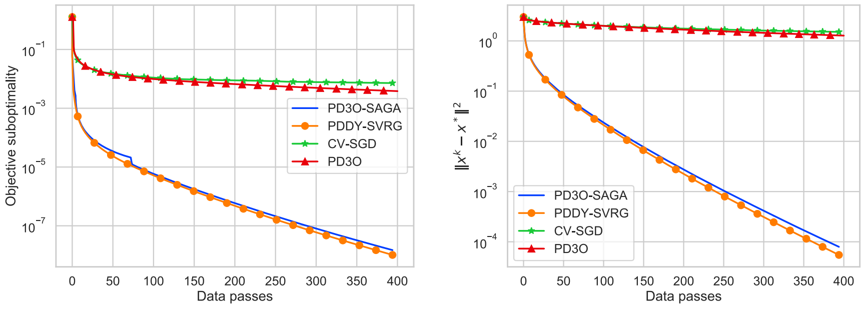

Fig. 1 Results for the PCA-Lasso experiment. Left: convergence w.r.t. the objective function; right: convergence in norm

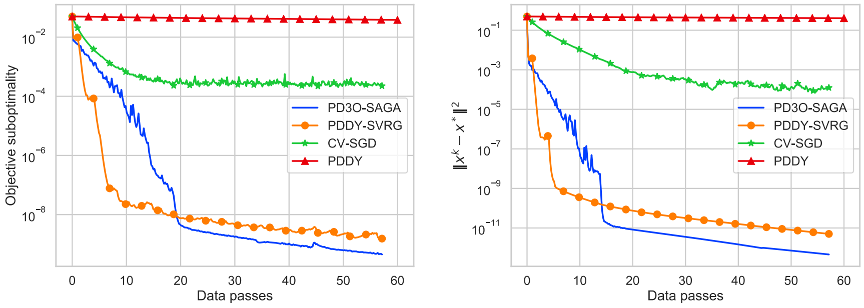

Fig. 2 Results for the MNIST experiment. Left: convergence w.r.t. the objective function; right: convergence in norm

(SGD), the VR estimators SAGA [29] and SVRG [40, 74, 79]. We used closed-form expressions for ν and tuned the stepsizes for all methods by running logarithmic grid search with factor 1.5 over multiples of $\frac{1}{\nu}$. We used a batch size of 16 for better parallelism in the stochastic estimators. For SGD, we used a small value of γ, such as $\frac{0.01}{\nu}$.

7.1 PCA-Lasso

In a recent work [71], the difficult PCA-based Lasso problem was considered: $\min_x \frac{1}{2}\|Wx - a\|^2 + \lambda\|x\|_1 + \lambda_1 \sum_{i=1}^m \|L_i x\|$, where $W \in \mathbb{R}^{n\times p}$, $a \in \mathbb{R}^n$, $\lambda, \lambda_1 > 0$ are given. We generated 10 matrices L_i randomly with standard normal i.i.d. entries, each with 20 rows. W and y were taken from the 'mushrooms' dataset in the libSVM base [14]. We chose $\lambda = \frac{\nu}{10n}$ and $\lambda_1 = \frac{2\nu}{nm}$, where ν is needed to compensate for the fact that we do not normalize the objective. The results are shown in Fig. 1. The advantage of using a VR stochastic gradient estimate is clear, with SAGA and SVRG being very similar.

7.2 MNIST with Overlapping Group Lasso

We consider the problem where F is the ℓ_2-regularized logistic loss and a group Lasso penalty. Given the data matrix $W \in \mathbb{R}^{n\times p}$ and vector of labels $a \in \{0,1\}^n$, $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) + \frac{\lambda}{2}\|x\|^2$, where $f_i(x) = -\big(a_i \log\big(h(w_i^\top x)\big) + (1-a_i)\log\big(1 - h(w_i^\top x)\big)\big)$, $\lambda = \frac{2\nu}{n}$, $w_i \in \mathbb{R}^p$ is the i-th row of W, and $h : t \to 1/(1+e^{-t})$. The nonsmooth regularizer is given by $\lambda_1 \sum_{j=1}^m \|x\|_{G_j}$, where $\lambda_1 = \frac{\nu}{5n}$, $G_j \subset \{1,\dots,p\}$ is a given subset of coordinates and $\|x\|_{G_j}$ is the ℓ_2-norm of the corresponding block of x. To apply splitting methods, we use $L = (I_{G_1}^\top, \dots, I_{G_m}^\top)^\top$, where I_{G_j} is the operator that takes $x \in \mathbb{R}^p$ and returns only the entries from block G_j. Then, we can use $H(y) = \lambda_1 \sum_{j=1}^m \|y\|_{G_j}$, which is separable in y and, thus, proximable. We use the MNIST dataset [48] of 70000 black and white 28 × 28 images. For each pixel, we add a group of pixels G_j adjacent to it, including the pixel itself. Since there are some border pixels, groups consist of 3, 4 or 5 coordinates, and there are 784 penalty terms in total. The results are shown in Fig. 2. Here SAGA is a bit better than SVRG.

7.3 Fused Lasso Experiment

In the Fused Lasso problem, we are given a feature matrix $W \in \mathbb{R}^{n\times p}$ and an output vector a, which define the least-squares function $F(x) = \frac{1}{2}\|Wx - a\|^2$. It is regularized with $\frac{\lambda}{2}\|x\|^2$ and $\lambda_1\|Dx\|_1$, where $\lambda = \frac{\nu}{n}$, $\lambda_1 = \frac{\nu}{10n}$ and $D \in \mathbb{R}^{(p-1)\times p}$ has entries $D_{i,i} = 1$, $D_{i,i+1} = -1$, for $i = 1,\dots,p-1$, and $D_{ij} = 0$ otherwise. We used again the ‘mushrooms’ dataset. The plots look very similar to the ones in Fig. 1, so we omit them.

8 Conclusion

We proposed a new primal–dual proximal splitting algorithm, the Primal–Dual Davis–Yin (PDDY) algorithm, to minimize a sum of three functions, one of which is composed with a linear operator. It is an alternative to the PD3O algorithm; they often perform similarly, but one or the other may be preferable for the problem at hand, depending on the implementation details. In particular, their memory requirements can be different. Furthermore, we proposed stochastic variants of both algorithms, studied their convergence rates, and showed by experiments that they can be much faster than their deterministic counterparts. We also showed that for linearly-constrained minimization of a strongly convex function, an instance of the stochastic PDDY algorithm, called LiCoSGD, converges linearly. We studied all algorithms within the unified framework of a stochastic generalization of Davis–Yin splitting for monotone inclusions. Our machinery opens the door to a promising class of new randomized proximal algorithms for large-scale optimization.

Appendix

A Lemmas A.1 and A.2

We state Lemma A.1 and Lemma A.2, which are used in the proofs of Theorem 5.1 and Theorem 5.2, respectively.

To simplify the notations, we use the following convention: when a set appears in an equation while a single element is expected, e.g. $\partial R(x^k)$, this means that the equation holds for some element in this nonempty set.

Lemma A.1 *Assume that F is μ_F-strongly convex, for some $\mu_F \geq 0$, and that $(g^k)_{k\in\mathbb{N}}$ satisfies Assumption 1. Then, the iterates of the Stochastic PD3O Algorithm satisfy*

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2+\kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 &\leq \|v^k-v^\star\|_P^2+\kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\ &-2\gamma(1-\gamma(\alpha+\kappa\delta))D_F(x^k,x^\star)-\gamma\mu_F\|x^k-x^\star\|^2\\ &-2\gamma\langle\partial R(x^k)-\partial R(x^\star),x^k-x^\star\rangle\\ &-2\gamma\mathbb{E}_k\langle\partial H^*(d^{k+1})-\partial H^*(d^\star),d^{k+1}-d^\star\rangle\\ &-\gamma^2\mathbb{E}_k\left\|P^{-1}A(u^{k+1})+P^{-1}B(z^k)-\left(P^{-1}A(u^\star)+P^{-1}B(z^\star)\right)\right\|_P^2.\end{aligned} \quad (22)$$

Proof Applying Lemma 3.2 for $\mathrm{DYS}(P^{-1}A, P^{-1}B, P^{-1}C)$ using the norm induced by P, we have

$$\begin{aligned}\|v^{k+1}-v^\star\|_P^2 &= \|v^k-v^\star\|_P^2-2\gamma\langle P^{-1}B(z^k)-P^{-1}B(z^\star),z^k-z^\star\rangle_P\\ &\quad -2\gamma\langle P^{-1}C(z^k)-P^{-1}C(z^\star),z^k-z^\star\rangle_P\\ &\quad +\gamma^2\|P^{-1}C(z^k)-P^{-1}C(z^\star)\|_P^2\\ &\quad -2\gamma\langle P^{-1}A(u^{k+1})-P^{-1}A(u^\star),u^{k+1}-u^\star\rangle_P\\ &\quad -\gamma^2\|P^{-1}A(u^{k+1})+P^{-1}B(z^k)-\left(P^{-1}A(u^\star)+P^{-1}B(z^\star)\right)\|_P^2\\ &= \|v^k-v^\star\|_P^2-2\gamma\langle B(z^k)-B(z^\star),z^k-z^\star\rangle\\ &\quad +\gamma^2\|P^{-1}C(z^k)-P^{-1}C(z^\star)\|_P^2\\ &\quad -2\gamma\langle C(z^k)-C(z^\star),z^k-z^\star\rangle-2\gamma\langle A(u^{k+1})-A(u^\star),u^{k+1}-u^\star\rangle\\ &\quad -\gamma^2\|P^{-1}A(u^{k+1})+P^{-1}B(z^k)-\left(P^{-1}A(u^\star)+P^{-1}B(z^\star)\right)\|_P^2.\end{aligned}$$

Using $A(u^{k+1})=\left(L^*d^{k+1},-Ls^{k+1}+\partial H^*(d^{k+1})\right)$, $B(z^k)=\left(\partial R(x^k),0\right)$, $C(z^k)=\left(g^{k+1},0\right)$ and $A(u^\star)=\left(L^*d^\star,-Ls^\star+\partial H^*(d^\star)\right)$, $B(z^\star)=\left(\partial R(x^\star),0\right)$, $C(z^\star)=\left(\nabla F(x^\star),0\right)$, we have

$$\begin{aligned}\|v^{k+1}-v^\star\|_P^2 &= \|v^k-v^\star\|_P^2-2\gamma\langle\partial R(x^k)-\partial R(x^\star),x^k-x^\star\rangle\\ &\quad +\gamma^2\|g^{k+1}-\nabla F(x^\star)\|^2\end{aligned}$$

$$
\begin{aligned}
&- 2\gamma\langle g^{k+1} - \nabla F(x^\star), x^k - x^\star\rangle - 2\gamma\langle \partial H^*(d^{k+1}) \\
&- \partial H^*(d^\star), d^{k+1} - d^\star\rangle \\
&- \gamma^2 \| P^{-1}A(u^{k+1}) + P^{-1}B(z^k) - \left(P^{-1}A(u^\star) + P^{-1}B(z^\star)\right) \|_P^2 .
\end{aligned}
$$

Taking conditional expectation w.r.t. $\mathscr{F}_k$ and using Assumption 1,

$$
\begin{aligned}
\mathbb{E}_k\|v^{k+1} - v^\star\|_P^2 \le \|v^k - v^\star\|_P^2 &- 2\gamma\langle \partial R(x^k) - \partial R(x^\star), x^k - x^\star\rangle \\
&- 2\gamma\langle \nabla F(x^k) - \nabla F(x^\star), x^k - x^\star\rangle \\
&- 2\gamma\mathbb{E}_k\langle \partial H^*(d^{k+1}) - \partial H^*(d^\star), d^{k+1} - d^\star\rangle \\
&+ \gamma^2\left(2\alpha D_F(x^k, x^\star) + \beta\sigma_k^2\right) \\
&- \gamma^2\mathbb{E}_k\| P^{-1}A(u^{k+1}) + P^{-1}B(z^k) \\
&- \left(P^{-1}A(u^\star) + P^{-1}B(z^\star)\right) \|_P^2 .
\end{aligned}
$$

Using strong convexity of F,

$$
\begin{aligned}
\mathbb{E}_k\|v^{k+1} - v^\star\|_P^2 \le \|v^k - v^\star\|_P^2 &- \gamma\mu_F\|x^k - x^\star\|^2 - 2\gamma D_F(x^k, x^\star) \\
&+ \gamma^2\left(2\alpha D_F(x^k, x^\star) + \beta\sigma_k^2\right) \\
&- 2\gamma\langle \partial R(x^k) - \partial R(x^\star), x^k - x^\star\rangle \\
&- 2\gamma\mathbb{E}_k\langle \partial H^*(d^{k+1}) - \partial H^*(d^\star), d^{k+1} - d^\star\rangle \\
&- \gamma^2\mathbb{E}_k\| P^{-1}A(u^{k+1}) + P^{-1}B(z^k) \\
&- \left(P^{-1}A(u^\star) + P^{-1}B(z^\star)\right) \|_P^2 .
\end{aligned}
$$

Using Assumption 1,

$$
\begin{aligned}
&\mathbb{E}_k\|v^{k+1} - v^\star\|_P^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 \le \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2 \\
&- \gamma\mu_F\|x^k - x^\star\|^2 \\
&- 2\gamma(1 - \gamma(\alpha + \kappa\delta))D_F(x^k, x^\star) - 2\gamma\langle \partial R(x^k) \\
&- \partial R(x^\star), x^k - x^\star\rangle \\
&- 2\gamma\mathbb{E}_k\langle \partial H^*(d^{k+1}) - \partial H^*(d^\star), d^{k+1} - d^\star\rangle \\
&- \gamma^2\mathbb{E}_k\left\| P^{-1}A(u^{k+1}) + P^{-1}B(z^k) \right. \\
&\left. - \left(P^{-1}A(u^\star) + P^{-1}B(z^\star)\right) \right\|_P^2 .
\end{aligned}
$$

□

Lemma A.2 *Suppose that $(g^k)_{k\in\mathbb{N}}$ satisfies Assumption 1. Then, the iterates of the Stochastic PDDY Algorithm satisfy*

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2+\kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 &\leq \|v^k-v^\star\|_P^2+\kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\ &-2\gamma(1-\gamma(\alpha+\kappa\delta))D_F(x^k,x^\star)-2\gamma\langle\partial H^*(y^k)\\ &-\partial H^*(y^\star),y^k-y^\star\rangle\\ &-2\gamma\mathbb{E}_k\langle\partial R(s^{k+1})-\partial R(s^\star),s^{k+1}-s^\star\rangle.\end{aligned}$$

Proof Applying Lemma 3.2 for DYS$(P^{-1}B,P^{-1}A,P^{-1}C)$ using the norm induced by P, we have

$$\begin{aligned}\|v^{k+1}-v^\star\|_P^2 &= \|v^k-v^\star\|_P^2-2\gamma\langle P^{-1}A(z^k)-P^{-1}A(z^\star),z^k-z^\star\rangle_P\\ &-2\gamma\langle P^{-1}C(z^k)-P^{-1}C(z^\star),z^k-z^\star\rangle_P\\ &+\gamma^2\|P^{-1}C(z^k)-P^{-1}C(z^\star)\|_P^2\\ &-2\gamma\langle P^{-1}B(u^{k+1})-P^{-1}B(u^\star),u^{k+1}-u^\star\rangle_P\\ &-\gamma^2\|P^{-1}B(u^{k+1})+P^{-1}A(z^k)-\left(P^{-1}B(u^\star)+P^{-1}A(z^\star)\right)\|_P^2\\ &=\|v^k-v^\star\|_P^2-2\gamma\langle A(z^k)-A(z^\star),z^k-z^\star\rangle-2\gamma\langle C(z^k)\\ &-C(z^\star),z^k-z^\star\rangle\\ &-2\gamma\langle B(u^{k+1})-B(u^\star),u^{k+1}-u^\star\rangle\\ &+\gamma^2\|P^{-1}C(z^k)-P^{-1}C(z^\star)\|_P^2\\ &-\gamma^2\|P^{-1}B(u^{k+1})+P^{-1}A(z^k)-\left(P^{-1}B(u^\star)+P^{-1}A(z^\star)\right)\|_P^2.\end{aligned}$$

Using $A(z^k)=\left(L^*y^k,-Lx^k+\partial H^*(y^k)\right)$, $B(u^{k+1})=\left(\partial R(s^{k+1}),0\right)$, $C(z^k)=\left(g^{k+1},0\right)$ and $A(z^\star)=\left(L^*y^\star,-Lx^\star+\partial H^*(y^\star)\right)$, $B(u^\star)=\left(\partial R(s^\star),0\right)$, $C(z^\star)=\left(\nabla F(x^\star),0\right)$, we have,

$$\begin{aligned}\|v^{k+1}-v^\star\|_P^2 &\leq \|v^k-v^\star\|_P^2-2\gamma\langle\partial H^*(y^k)-\partial H^*(y^\star),y^k-y^\star\rangle\\ &+\gamma^2\|g^{k+1}-\nabla F(x^\star)\|^2\\ &-2\gamma\langle g^{k+1}-\nabla F(x^\star),x^k-x^\star\rangle-2\gamma\langle\partial R(s^{k+1})\\ &-\partial R(s^\star),s^{k+1}-s^\star\rangle.\end{aligned}$$

Applying the conditional expectation w.r.t. $\mathscr{F}_k$ and using Assumption 1,

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2 &\leq \|v^k-v^\star\|_P^2-2\gamma\langle\partial H^*(y^k)-\partial H^*(y^\star),y^k-y^\star\rangle\\ &-2\gamma\langle\nabla F(x^k)-\nabla F(x^\star),x^k-x^\star\rangle\\ &+\gamma^2\left(2\alpha D_F(x^k,x^\star)+\beta\sigma_k^2\right)\end{aligned}$$

$$- 2\gamma \mathbb{E}_k \langle \partial R(s^{k+1}) - \partial R(s^\star), s^{k+1} - s^\star \rangle.$$

Using the convexity of F,

$$\begin{aligned}\mathbb{E}_k \|v^{k+1} - v^\star\|_P^2 \le \|v^k - v^\star\|_P^2 &- 2\gamma \langle \partial H^*(y^k) - \partial H^*(y^\star), y^k - y^\star \rangle \\ &- 2\gamma D_F(x^k, x^\star) \\ &- 2\gamma \mathbb{E}_k \langle \partial R(s^{k+1}) - \partial R(s^\star), s^{k+1} - s^\star \rangle \\ &+ \gamma^2 \left(2\alpha D_F(x^k, x^\star) + \beta \sigma_k^2\right).\end{aligned}$$

Using Assumption 1,

$$\begin{aligned}&\mathbb{E}_k \|v^{k+1} - v^\star\|_P^2 + \kappa\gamma^2 \mathbb{E}_k \sigma_{k+1}^2 \le \|v^k - v^\star\|_P^2 + \kappa\gamma^2 \left(1 - \rho + \frac{\beta}{\kappa}\right) \sigma_k^2 \\ &\quad - 2\gamma(1 - \gamma(\alpha + \kappa\delta)) D_F(x^k, x^\star) \\ &\quad - 2\gamma \langle \partial H^*(y^k) - \partial H^*(y^\star), y^k - y^\star \rangle \\ &\quad - 2\gamma \mathbb{E}_k \langle \partial R(s^{k+1}) - \partial R(s^\star), s^{k+1} - s^\star \rangle.\end{aligned}$$

□

B Linear Convergence Results

In this section, we provide linear convergence results for the stochastic PD3O and the stochastic PDDY algorithms, in addition to Theorem 6.2. For an operator splitting method like DYS$(\tilde{A}, \tilde{B}, \tilde{C})$ to converge linearly, it is necessary that $\tilde{A} + \tilde{B} + \tilde{C}$ is strongly monotone. But this is not sufficient, and in general,

to converge linearly, DYS$(\tilde{A}, \tilde{B}, \tilde{C})$ requires the stronger assumption that $\tilde{A}$ or $\tilde{B}$ or $\tilde{C}$ is strongly monotone, and in addition that $\tilde{A}$ or $\tilde{B}$ is cocoercive [28]. The PDDY algorithm is equivalent to DYS$(P^{-1}B, P^{-1}A, P^{-1}C)$ and the PD3O algorithm is equivalent to DYS$(P^{-1}A, P^{-1}B, P^{-1}C)$, see Sect. 4. However, $P^{-1}A$, $P^{-1}B$ and $P^{-1}C$ are not strongly monotone. In spite of this, we will prove linear convergence of the (stochastic) PDDY and PD3O algorithms.

Thus, for both algorithms, we will make the assumption that $P^{-1}A+P^{-1}B+P^{-1}C$ is strongly monotone. This is equivalent to assuming that $M = A+B+C$ is strongly monotone; that is, that $F + R$ is strongly convex and H is smooth. For instance, the Chambolle–Pock algorithm [11, 13], which is a particular case of the PD3O and the PDDY algorithms, requires R strongly convex and H smooth to converge linearly, in general. In fact, for primal–dual algorithms to converge linearly on Problem (1), for any L, it seems unavoidable that $F + R$ is strongly convex and that the dual term H^* is strongly convex too, because the algorithm needs to be contractive in both the primal and the dual spaces. This means that H must be smooth. We can remark that if H is smooth, it is tempting to use its gradient instead of its proximity operator. We can then use the proximal gradient algorithm to solve Problem (1) with $\nabla(F + H \circ L)(x) = \nabla F(x) + L^*\nabla H(Lx)$. However, in practice, it is often faster to use the proximity operator instead of the gradient, see a recent analysis of this topic in [18].

For the PD3O algorithm, we will add a cocoercivity assumption, as suggested by the general linear convergence theory of DYS. More precisely, we will assume that R is smooth, so that $P^{-1}B$ is cocoercive. Our result on the PD3O is therefore an extension of [77][Theorem 3] to the stochastic setting. For the PDDY algorithm, this assumption is not needed to prove linear convergence, which is an advantage over the PD3O algorithm.

We denote by $\|\cdot\|_{\gamma,\tau}$ the norm induced by $\frac{\gamma}{\tau}I - \gamma^2 LL^*$ on $\mathcal{Y}$.

Theorem B.1 *(Linear convergence of the Stochastic PD3O Algorithm) Suppose that Assumption 1 holds. Suppose that H is $1/\mu_{H^*}$-smooth, for some $\mu_{H^*} > 0$, F is μ_F-strongly convex, for some $\mu_F \geq 0$, and R is μ_R-strongly convex, for some $\mu_R \geq 0$, with $\mu := \mu_F + 2\mu_R > 0$. Also, suppose that R is λ-smooth, for some $\lambda > 0$. Suppose that the parameters $\gamma > 0$ and $\tau > 0$ satisfy $\gamma \leq 1/(\alpha + \kappa\delta)$, for some $\kappa > \beta/\rho$, and $\gamma\tau\|L\|^2 < 1$. Define, for every $k \in \mathbb{N}$,*

$$V^k := \|p^k - p^\star\|^2 + (1 + 2\tau\mu_{H^*})\,\|y^k - y^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\sigma_k^2, \tag{23}$$

and

$$r := \max\left(1 - \frac{\gamma\mu}{(1+\gamma\lambda)^2}, \left(1 - \rho + \frac{\beta}{\kappa}\right), \frac{1}{1 + 2\tau\mu_{H^*}}\right). \tag{24}$$

Then, for every $k \in \mathbb{N}$, $\mathbb{E}V^k \leq r^k V^0$.

Proof We first use Lemma A.1 along with the strong convexity of R, H^*. Note that $y^k = q^k$ and therefore $q^{k+1} = q^k + d^{k+1} - q^k = d^{k+1}$. We have

$$\begin{aligned}
&\mathbb{E}_k\|p^{k+1} - p^\star\|^2 + \mathbb{E}_k\|q^{k+1} - q^\star\|^2_{\gamma,\tau} + 2\gamma\mu_{H^*}\mathbb{E}_k\|q^{k+1} - q^\star\|^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2\\
&\leq \|p^k - p^\star\|^2 + \|q^k - q^\star\|^2_{\gamma,\tau} - \gamma\mu\|x^k - x^\star\|^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2\\
&\quad - 2\gamma(1 - \gamma(\alpha + \kappa\delta))D_F(x^k, x^\star).
\end{aligned}$$

Noting that for every $q \in \mathcal{Y}$, $\|q\|^2_{\gamma,\tau} = \frac{\gamma}{\tau}\|q\|^2 - \gamma^2\|L^*q\|^2 \leq \frac{\gamma}{\tau}\|q\|^2$, and taking $\gamma \leq 1/(\alpha + \kappa\delta)$, we have

$$\begin{aligned}
&\mathbb{E}_k\|p^{k+1} - p^\star\|^2 + (1 + 2\tau\mu_{H^*})\,\mathbb{E}_k\|q^{k+1} - q^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2\\
&\leq \|p^k - p^\star\|^2 + \|q^k - q^\star\|^2_{\gamma,\tau} - \gamma\mu\|x^k - x^\star\|^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2.
\end{aligned}$$

Finally, since R is λ-smooth, $\|p^k - p^\star\|^2 \leq (1 + 2\gamma\lambda + \gamma^2\lambda^2)\|x^k - x^\star\|^2$. Indeed, in this case, applying Lemma 3.2 with $\tilde{A} = 0$, $\tilde{C} = 0$ and $\tilde{B} = \nabla R$, we obtain that if

$x^k = \operatorname{prox}_{\gamma R}(p^k)$ and $x^\star = \operatorname{prox}_{\gamma R}(p^\star)$, then

$$\begin{aligned}\|x^k - x^\star\|^2 = \|p^k - p^\star\|^2 - 2\gamma\langle \nabla R(x^k) - \nabla R(x^\star), x^k - x^\star\rangle \\ - \gamma^2\|\nabla R(x^k) - \nabla R(x^\star)\|^2 \\ \geq \|p^k - p^\star\|^2 - 2\gamma\lambda\|x^k - x^\star\|^2 - \gamma^2\lambda^2\|x^k - x^\star\|^2.\end{aligned}$$

Hence,

$$\begin{aligned}&\mathbb{E}_k\|p^{k+1} - p^\star\|^2 + (1 + 2\tau\mu_{H^*})\,\mathbb{E}_k\|q^{k+1} - q^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}_k\sigma^2_{k+1} \\ &\leq \|p^k - p^\star\|^2 + \|q^k - q^\star\|^2_{\gamma,\tau} - \frac{\gamma\mu}{(1+\gamma\lambda)^2}\|p^k - p^\star\|^2 \\ &+ \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2.\end{aligned}$$

Thus, by setting V^k as in (23) and r as in (24), we have $\mathbb{E}_k V^{k+1} \leq r V^k$. □

Thus, under smoothness and strong convexity assumptions, Theorem B.1 implies linear convergence of the dual variable y^k to $y^\star$, with convergence rate given by r. Since $\|x^k - x^\star\| \leq \|p^k - p^\star\|$, it also implies linear convergence of the variable x^k to $x^\star$, with same rate.

If $g^{k+1} = \nabla F(x^k)$, the Stochastic PD3O Algorithm reverts to the PD3O Algorithm and Theorem B.1 provides a convergence rate similar to Theorem 3 in [77]. In this case, by taking $\kappa = 1$, we obtain

$$r = \max\left(1 - \gamma\frac{\mu_F + 2\mu_R}{(1+\gamma\lambda)^2}, \frac{1}{1 + 2\tau\mu_{H^*}}\right),$$

whereas Theorem 3 in [77] provides the rate

$$\max\left(1 - \gamma\frac{2(\mu_F + \mu_R) - \gamma\alpha\mu_F}{(1+\gamma\lambda)^2}, \frac{1}{1 + 2\tau\mu_{H^*}}\right)$$

(the reader might not recognize the rate given in Theorem 3 of [77] because of some typos in Eqn. 39 of [77]).

Theorem B.2 *(Linear convergence of the Stochastic PDDY Algorithm) Suppose that Assumption 1 holds. Also, suppose that H is $1/\mu_{H^*}$-smooth and R is μ_R-strongly convex, for some $\mu_R > 0$ and $\mu_{H^*} > 0$. Suppose that the parameters $\gamma > 0$ and $\tau > 0$ satisfy $\gamma \leq 1/(\alpha + \kappa\delta)$, for some $\kappa > \beta/\rho$, $\gamma\tau\|L\|^2 < 1$, and $\gamma^2 \leq \frac{\mu_{H^*}}{\|L\|^2\mu_R}$. Define $\eta := 2\left(\mu_{H^*} - \gamma^2\|L\|^2\mu_R\right) \geq 0$ and, for every $k \in \mathbb{N}$,*

$$V^k := (1 + \gamma\mu_R)\|p^k - p^\star\|^2 + (1 + \tau\eta)\|y^k - y^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\sigma_k^2, \tag{25}$$

and

$$r := \max\left(\frac{1}{1+\gamma\mu_R}, 1-\rho+\frac{\beta}{\kappa}, \frac{1}{1+\tau\eta}\right). \tag{26}$$

Then, for every $k \in \mathbb{N}$, $\mathbb{E}V^k \le r^k V^0$.

Proof We first use Lemma A.2 along with the strong convexity of R and H^*. Note that $y^k = q^{k+1}$. We have

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 \le \|v^k - v^\star\|_P^2 &+ \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\ &- 2\gamma\mu_{H^*}\mathbb{E}_k\|q^{k+1}-q^\star\|^2\\ &- 2\gamma\mu_R\mathbb{E}_k\|s^{k+1}-s^\star\|^2.\end{aligned}$$

Note that $s^{k+1} = p^{k+1} - \gamma L^* y^k$. Therefore, $s^{k+1} - s^\star = (p^{k+1} - p^\star) - \gamma L^*(y^k - y^\star)$. Using Young's inequality $-\|a+b\|^2 \le -\frac{1}{2}\|a\|^2 + \|b\|^2$, we have $-\mathbb{E}_k\|s^{k+1}-s^\star\|^2 \le -\frac{1}{2}\mathbb{E}_k\|p^{k+1}-p^\star\|^2 + \gamma^2\|L\|^2\mathbb{E}_k\|q^{k+1}-q^\star\|^2$. Hence, using $\tau\|q\|_{\gamma,\tau}^2 \le \gamma\|q\|^2$,

$$\begin{aligned}\mathbb{E}_k\|v^{k+1}-v^\star\|_P^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 &\le \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\ &\quad - 2\gamma\left(\mu_{H^*} - \gamma^2\|L\|^2\mu_R\right)\mathbb{E}_k\|q^{k+1}-q^\star\|^2\\ &\quad - \gamma\mu_R\mathbb{E}_k\|p^{k+1}-p^\star\|^2\\ &\le \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2\\ &\quad - 2\tau\mathbb{E}_k\|q^{k+1}-q^\star\|_{\gamma,\tau}^2\left(\mu_{H^*} - \gamma^2\|L\|^2\mu_R\right)\\ &\quad - \gamma\mu_R\mathbb{E}_k\|p^{k+1}-p^\star\|^2.\end{aligned}$$

Set $\eta := 2\left(\mu_{H^*} - \gamma^2\|L\|^2\mu_R\right) \ge 0$. Then,

$$\begin{aligned}&(1+\gamma\mu_R)\mathbb{E}_k\|p^{k+1}-p^\star\|^2 + (1+\tau\eta)\mathbb{E}_k\|q^{k+1}-q^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2\\ &\quad\le \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1-\rho+\frac{\beta}{\kappa}\right)\sigma_k^2.\end{aligned}$$

Thus, by setting V^k as in (25) and r as in (26), we have $\mathbb{E}_k V^{k+1} \le rV^k$. □

C PriLiCoSGD and Application to Decentralized Optimization

Algorithm 5 *

```
PriLiCoSGD (new)
Input: $x^0 \in \mathcal{X}$, $a^0 \in \mathrm{ran}(W)$, $\gamma > 0$, $\tau > 0$
for $k = 0, 1, 2, \dots$ do
    $t^{k+1} = x^k - \gamma g^{k+1}$
    $a^{k+1} = a^k + \tau W(t^{k+1} - \gamma a^k) - \tau c$
    $x^{k+1} = t^{k+1} - \gamma a^{k+1}$
end for
```

Algorithm 6 *

```
DESTROY (new)
Input: $x_i^0 \in \mathcal{X}$ and $a_i^0 \in \mathcal{X}$, $\forall i \in V$, such that
    $\sum_{i \in V} a_i^0 = 0$, $\gamma > 0$, $\tau > 0$
for $k = 0, 1, 2, \dots$ do
    for all $i \in V$ in parallel do
        $t_i^{k+1} = x_i^k - \gamma g_i^{k+1}$
        $a_i^{k+1} = (1 - \tau\gamma \widehat{W}_{i,i}) a_i^k + \tau \widehat{W}_{i,i} t_i^{k+1}$
        $+ \tau \sum_{j \neq i : \{i,j\} \in V} \widehat{W}_{i,j} (t_j^{k+1} - \gamma a_j^k)$
        $x_i^{k+1} = t_i^{k+1} - \gamma a_i^{k+1}$.
    end for
end for
```

In decentralized optimization, a network of computing agents aims at jointly minimizing an objective function by performing local computations and exchanging information along the edges [1, 46, 66, 67]. It is a particular case of linearly constrained optimization, as detailed below.

First, let us set $W := L^*L$ and $c := L^*b$. Replacing the variable y^k by the variable $a^k := L^*y^k$ in LiCoSGD, we can write the algorithm using W and c instead of L, L^* and b, with primal variables in $\mathcal{X}$ only. This yields the new algorithm PriLiCoSGD, shown above, to minimize $F(x)$ subject to $Wx = c$. The convergence results for LiCoSGD apply to PriLiCoSGD, with $(a^k)_{k \in \mathbb{N}}$ converging to $a^\star = -\nabla F(x^\star)$.

We can apply PriLiCoSGD to decentralized optimization as follows. Consider a connected undirected graph $G = (V, E)$, where $V = \{1, \dots, N\}$ is the set of nodes and E the set of edges. Consider a family $(f_i)_{i \in V}$ of μ-strongly convex and ν-smooth functions f_i, for some $\mu \geq 0$ and $\nu > 0$. The problem is:

$$\min_{x \in \mathcal{X}} \sum_{i \in V} f_i(x). \tag{27}$$

Consider a gossip matrix of the graph G; that is, a $N \times N$ symmetric positive semidefinite matrix $\widehat{W} = (\widehat{W}_{i,j})_{i,j \in V}$, such that $\ker(\widehat{W}) = \mathrm{span}([1 \cdots 1]^{\mathrm{T}})$ and $\widehat{W}_{i,j} \neq 0$ if and only if $i = j$ or $\{i, j\} \in E$ is an edge of the graph. $\widehat{W}$ can be the Laplacian matrix of G, for instance. Set $W := \widehat{W} \otimes I$, where $\otimes$ is the Kronecker product; then decentralized communication in the network G is modeled by an application of the positive self-adjoint linear operator W on $\mathcal{X}^V$. Moreover, $W(x_1, \dots, x_N) = 0$ if and only if $x_1 = \dots = x_N$. Therefore, Problem (27) is equivalent to the lifted problem

$$\min_{\tilde{x} \in \mathcal{X}^V} F(\tilde{x}) \quad \text{such that} \quad W\tilde{x} = 0, \tag{28}$$

where for every $\tilde{x} = (x_1, \dots, x_N) \in \mathcal{X}^V$, $F(\tilde{x}) = \sum_{i=1}^N f_i(x_i)$. Let us apply PriLiCoSGD to Problem (28); we obtain the Decentralized Stochastic Optimization Algorithm (DESTROY). It generates the sequence $(\tilde{x}^k)_{k \in \mathbb{N}}$, where $\tilde{x}^k =$

$(x_1^k,\dots,x_N^k) \in \mathcal{X}^V$. The update of each x_i^k consists in evaluating g_i^{k+1}, an estimate of $\nabla f_i(x_i^k)$ satisfying Assumption 1,

and communication steps involving x_j^k, for every neighbor j of i. For instance, the variance-reduced estimator g_i^k can be the loopless SVRG estimator seen in Proposition 5.1, when f_i is itself a sum of functions, or a compressed version of ∇f_i [4, 50, 65, 75].

As an application of the convergence results for LiCoSGD, we obtain the following results for DESTROY. Theorem 4.1 becomes:

Theorem C.1 *(Convergence of DESTROY, deterministic case $g_i^{k+1} = \nabla f_i(x_i^k)$) Suppose that $\gamma \in (0, 2/\nu)$ and that $\tau\gamma\|\widehat{W}\| < 1$. Then, in DESTROY, each $(x_i^k)_{k\in\mathbb{N}}$ converges to the same solution $x^\star$ to the problem (27) and each $(a_i^k)_{k\in\mathbb{N}}$ converges to $a_i^\star = -\nabla f_i(x^\star)$.*

Theorem 6.1 can be applied to the stochastic case, stating $\mathcal{O}(1/k)$ convergence of the Lagrangian gap, by setting $\mathcal{Y} = \mathcal{X}$ and $L = L^* = W^{1/2}$. Similarly, Theorem 6.2 yields linear convergence of DESTROY in the strongly convex case $\mu > 0$, with L^*L replaced by W and $\|L\|^2$ replaced by $\|W\| = \|\widehat{W}\|$. In particular, in the deterministic case, with $\gamma = 1/\nu$ and $\tau\gamma = \aleph/\|W\|$ for some fixed $\aleph \in (0, 1)$, ε-accuracy is reached after $\mathcal{O}\Big(\max\big(\frac{\nu}{\mu}, \frac{\|W\|}{\omega(W)}\big)\log\big(\frac{1}{\varepsilon}\big)\Big)$ iterations. This rate is better or equivalent to the one of recently proposed decentralized algorithms, like EXTRA, DIGing, NIDS, NEXT, Harness, Exact Diffusion, see Table 1 of [76], [49][Theorem 1] and [1]. With a stochastic gradient, the rate of our algorithm is also better than [53][Equation 99].

In follow-up papers, the authors used Nesterov acceleration to propose accelerated versions of DESTROY [46] and PriLiCoSGD [64].

References

1. Alghunaim, S.A., Ryu, E.K., Yuan, K., Sayed, A.H.: Decentralized proximal gradient algorithms with linear convergence rates. IEEE Trans. Autom. Control **66**(6), 2787–2794 (2021)
2. Alotaibi, A., Combettes, P.L., Shahzad, N.: Solving coupled composite monotone inclusions by successive Fejér approximations of their Kuhn-Tucker set. SIAM J. Optim. **24**(4), 2076–2095 (2014)
3. Bach, F., Jenatton, R., Mairal, J., Obozinski, G.: Optimization with sparsity-inducing penalties. Found. Trends Mach. Learn. **4**(1), 1–106 (2012)
4. Basu, D., Data, D., Karakus, C., Diggavi, S.N.: Qsparse-Local-SGD: distributed SGD with quantization, sparsification, and local computations. IEEE J. Select. Areas Inform. Theor. **1**(1), 217–226 (2020)
5. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. Springer, New York (2017)
6. Beck, A.: First-Order Methods in Optimization. SIAM, MOS-SIAM Series on Optimization (2017)
7. Boţ, R.I., Csetnek, E.R., Hendrich, C.: Recent developments on primal-dual splitting methods with applications to convex minimization. In: Pardalos, P.M., Rassias, T.M. (eds.) Mathematics Without Boundaries: Surveys in Interdisciplinary Research, pp. 57–99. Springer, New York (2014)
8. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
9. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press (2004)
10. Bredies, K., Kunisch, K., Pock, T.: Total generalized variation. SIAM J. Imaging Sci. **3**(3), 492–526 (2010)

11. Chambolle, A., Pock, T.: A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
12. Chambolle, A., Pock, T.: An introduction to continuous optimization for imaging. Acta Numer. **25**, 161–319 (2016)
13. Chambolle, A., Pock, T.: On the ergodic convergence rates of a first-order primal-dual algorithm. Math. Program. **159**(1–2), 253–287 (2016)
14. Chang, C.C., Lin, C.J.: LibSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. (TIST) **2**(3), 27 (2011)
15. Chen, P., Huang, J., Zhang, X.: A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration. Inverse Probl. **29**(2), 025011 (2013)
16. Combettes, P.L., Condat, L., Pesquet, J.C., Vũ, B.C.: A forward–backward view of some primal–dual optimization methods in image recovery. In: Proc. of IEEE ICIP. Paris, France (2014)
17. Combettes, P.L., Eckstein, J.: Asynchronous block-iterative primal-dual decomposition methods for monotone inclusions. Math. Program. **168**(1–2), 645–672 (2018)
18. Combettes, P.L., Glaudin, L.E.: Proximal activation of smooth functions in splitting algorithms for convex image recovery. SIAM J. Imaging Sci. **12**(4), 1905–1935 (2019)
19. Combettes, P.L., Pesquet, J.C.: Proximal splitting methods in signal processing. In: Bauschke, H.H., Burachik, R., Combettes, P.L., Elser, V., Luke, D.R., Wolkowicz, H. (eds.) Fixed-Point Algorithms for Inverse Problems in Science and Engineering. Springer-Verlag, New York (2010)
20. Combettes, P.L., Pesquet, J.C.: Primal-dual splitting algorithm for solving inclusions with mixtures of composite, Lipschitzian, and parallel-sum type monotone operators. Set-Val. Var. Anal. **20**(2), 307–330 (2012)
21. Combettes, P.L., Pesquet, J.C.: Fixed point strategies in data science. IEEE Trans. Signal Process. **69**, 3878–3905 (2021)
22. Condat, L.: A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**(2), 460–479 (2013)
23. Condat, L.: A generic proximal algorithm for convex optimization–application to total variation minimization. IEEE Signal Process. Lett. **21**(8), 1054–1057 (2014)
24. Condat, L.: Discrete total variation: new definition and minimization. SIAM J. Imaging Sci. **10**(3), 1258–1290 (2017)
25. Condat, L., Kitahara, D., Contreras, A., Hirabayashi, A.: Proximal splitting algorithms for convex optimization: a tour of recent advances, with new twists. SIAM Review . To appear (2022)
26. Condat, L., Malinovsky, G., Richtárik, P.: Distributed proximal splitting algorithms with rates and acceleration. Front. Signal Process. (2022). https://doi.org/10.3389/frsip.2021.776825
27. Couprie, C., Grady, L., Najman, L., Pesquet, J.C., Talbot, H.: Dual constrained TV-based regularization on graphs. SIAM J. Imaging Sci. **6**(3), 1246–1273 (2013)
28. Davis, D., Yin, W.: A three-operator splitting scheme and its optimization applications. Set-Val. Var. Anal. **25**, 829–858 (2017)
29. Defazio, A., Bach, F., Lacoste-Julien, S.: Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Weinberger (eds.) Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc. (2014)
30. Drori, Y., Sabach, S., Teboulle, M.: A simple algorithm for a class of nonsmooth convex concave saddle-point problems. Oper. Res. Lett. **43**(2), 209–214 (2015)
31. Duran, J., Moeller, M., Sbert, C., Cremers, D.: Collaborative total variation: A general framework for vectorial TV models. SIAM J. Imaging Sci. **9**(1), 116–151 (2016)
32. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**, 293–318 (1992)
33. Eckstein, J., Svaiter, B.F.: A family of projective splitting methods for the sum of two maximal monotone operators. Math. Program. **111**(1), 173–199 (2008)
34. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
35. Glowinski, R., Marrocco, A.: Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. Revue Française d'Automatique, Informatique et Recherche Opérationnelle **9**, 41–76 (1975)

36. Gorbunov, E., Hanzely, F., Richtárik, P.: A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In: S. Chiappa, R. Calandra (eds.) Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS), vol. PMLR 108, pp. 680–690 (2020)
37. Gower, R.M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., Richtárik, P.: SGD: General analysis and improved rates. In: K. Chaudhuri, R. Salakhutdinov (eds.) Proc. of 36th Int. Conf. Machine Learning (ICML), vol. PMLR 97, pp. 5200–5209 (2019)
38. Gower, R.M., Schmidt, M., Bach, F., Richtárik, P.: Variance-reduced methods for machine learning. Proc. IEEE **108**(11), 1968–1983 (2020)
39. Hofmann, T., Lucchi, A., Lacoste-Julien, S., McWilliams, B.: Variance reduced stochastic gradient descent with neighbors. In: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 28, pp. 2305–2313. Curran Associates, Inc. (2015)
40. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) Advances in Neural Information Processing Systems, vol. 26, pp. 315–323. Curran Associates, Inc. (2013)
41. Johnstone, P.R., Eckstein, J.: Convergence rates for projective splitting. SIAM J. Optim. **29**(3), 1931–1957 (2019)
42. Johnstone, P.R., Eckstein, J.: Single-forward-step projective splitting: exploiting cocoercivity. Comput. Optim. Appl. **78**(1), 125–166 (2021)
43. Johnstone, P.R., Eckstein, J.: Projective splitting with forward steps. Math. Program. **191**, 631–670 (2022)
44. Komodakis, N., Pesquet, J.C.: Playing with duality: an overview of recent primal-dual approaches for solving large-scale optimization problems. IEEE Signal Process. Mag. **32**(6), 31–54 (2015)
45. Kovalev, D., Horváth, S., Richtárik, P.: Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In: A. Kontorovich, G. Neu (eds.) Proc. of Int. Conf. Algo. Learn. Theory (ALT), vol. PMLR 117, pp. 451–467 (2020)
46. Kovalev, D., Salim, A., Richtárik, P.: Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 18342–18352. Curran Associates, Inc. (2020)
47. Lan, G.: First-order and Stochastic Optimization Methods for Machine Learning. Springer Cham (2020)
48. LeCun, Y., Cortes, C.: MNIST handwritten digit database (2010). http://yann.lecun.com/exdb/mnist/
49. Li, H., Lin, Z.: Revisiting EXTRA for smooth distributed optimization. SIAM J. Optim. **30**(3), 1795–1821 (2020)
50. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Process. Mag. **3**(37), 50–60 (2020)
51. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
52. Loris, I., Verhoeven, C.: On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty. Inverse Probl. **27**(12) (2011)
53. Mokhtari, A., Ribeiro, A.: DSA: Decentralized double stochastic averaging gradient algorithm. J. Mach. Learn. Res. **17**(1), 2165–2199 (2016)
54. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer (2018)
55. O'Connor, D., Vandenberghe, L.: On the equivalence of the primal-dual hybrid gradient method and Douglas-Rachford splitting. Math. Program. **79**, 85–108 (2020)
56. Palomar, D.P., Eldar, Y.C. (eds.): Convex Optimization in Signal Processing and Communications. Cambridge University Press (2009)
57. Parikh, N., Boyd, S.: Proximal algorithms. Found. Trends Optim. **3**(1), 127–239 (2014)
58. Pedregosa, F., Fatras, K., Casotto, M.: Proximal splitting meets variance reduction. In: K. Chaudhuri, M. Sugiyama (eds.) Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS), vol. PMLR 89, pp. 1–10 (2019)
59. Polson, N.G., Scott, J.G., Willard, B.T.: Proximal algorithms in statistics and machine learning. Statist. Sci. **30**(4), 559–581 (2015)
60. Pustelnik, N., Condat, L.: Proximity operator of a sum of functions; application to depth map estimation. IEEE Signal Process. Lett. **24**(12), 1827–1831 (2017)
61. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**(1–4), 259–268 (1992)

62. Ryu, E.K.: Uniqueness of DRS as the 2 operator resolvent-splitting and impossibility of 3 operator resolvent-splitting. Math. Program. **182**, 233–273 (2020)
63. Salim, A., Bianchi, P., Hachem, W.: Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. IEEE Trans. Automat. Contr. **64**(5), 1832–1847 (2019)
64. Salim, A., Condat, L., Kovalev, D., Richtárik, P.: An optimal algorithm for strongly convex minimization under affine constraints. In: G. Camps-Valls, F.J.R. Ruiz, I. Valera (eds.) Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS), vol. PMLR 151, pp. 4482–4498 (2022)
65. Sattler, F., Wiedemann, S., K.-R. Müller, Samek, W.: Robust and communication-efficient federated learning from non-i.i.d. data. IEEE Trans. Neural Networks and Learning Systems **31**(9), 3400–3413 (2020)
66. Scaman, K., Bach, F., Bubeck, S., Lee, Y.T., Massoulié, L.: Optimal algorithms for smooth and strongly convex distributed optimization in networks. In: D. Precup, Y.W. Teh (eds.) Proc. of 34th Int. Conf. Machine Learning (ICML), vol. PMLR 70, pp. 3027–3036 (2017)
67. Shi, W., Ling, Q., Wu, G., Yin, W.: EXTRA: An exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**(2), 944–966 (2015)
68. Starck, J.L., Murtagh, F., Fadili, J.: Sparse Image and Signal Processing: Wavelets, Curvelets. Cambridge University Press, Morphological Diversity (2010)
69. Stathopoulos, G., Shukla, H., Szucs, A., Pu, Y., Jones, C.N.: Operator splitting methods in control. Found. Trends Syst. Control **3**(3), 249–362 (2016)
70. Svaiter, B.F.: On weak convergence of the Douglas-Rachford method. SIAM J. Control. Optim. **49**(1), 280–287 (2011)
71. Tay, J.K., Friedman, J., Tibshirani, R.: Principal component-guided sparse regression. Can. J. Stat. **49**, 1222–1257 (2021)
72. Vũ, B.C.: A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**(3), 667–681 (2013)
73. Wright, S.J.: Coordinate descent algorithms. Math. Program. **151**, 3–34 (2015)
74. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)
75. Xu, H., Ho, C.Y., Abdelmoniem, A.M., Dutta, A., Bergou, E.H., Karatsenidis, K., Canini, M., Kalnis, P.: GRACE: A compressed communication framework for distributed machine learning. In: Proc. of 41st IEEE Int. Conf. Distributed Computing Systems (ICDCS), pp. 561–572 (2021)
76. Xu, J., Tian, Y., Sun, Y., Scutari, G.: Distributed algorithms for composite optimization: unified framework and convergence analysis. IEEE Trans. Signal Process. **69**, 3555–3570 (2021)
77. Yan, M.: A new Primal-Dual algorithm for minimizing the sum of three functions with a linear operator. J. Sci. Comput. **76**(3), 1698–1717 (2018)
78. Yurtsever, A., Vu, B.C., Cevher, V.: Stochastic three-composite convex minimization. In: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 4329–4337. Curran Associates, Inc. (2016)
79. Zhang, L., Mahdavi, M., Jin, R.: Linear convergence with condition number independent access of full gradients. In: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (eds.) Advances in Neural Information Processing Systems, vol. 26, pp. 980–988. Curran Associates, Inc. (2013)
80. Zhao, R., Cevher, V.: Stochastic three-composite convex minimization with a linear operator. In: A. Storkey, F. Perez-Cruz (eds.) Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS), vol. PMLR 84, pp. 765–774 (2018)