OPT2020: 12th Annual Workshop on Optimization for Machine Learning

Dualize, Split, Randomize: Fast Nonsmooth Optimization Algorithms

Adil Salim ADIL.SALIM@KAUST.EDU.SA
Laurent Condat LAURENT.CONDAT@KAUST.EDU.SA
Konstantin Mishchenko KONSTANTIN.MISHCHENKO@KAUST.EDU.SA
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
KAUST, Thuwal, Saudi Arabia

Abstract

We consider[1] the task of 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. First, we propose a new primal dual algorithm called PDDY to solve such problem. PDDY can be seen as an instance of Davis–Yin splitting involving operators which are monotone under a new metric depending on L. This representation of PDDY eases the non asymptotic analysis of PDDY: it enables us to prove its sublinear convergence (resp. linear convergence if strong convexity is involved), even when a variance reduced stochastic gradient of F is used instead of the full gradient. Moreover, we surprisingly obtain as a special case an algorithm for the minimization of a strongly convex F under affine constraints $Lx = b$, linearly converging without projecting onto the constraints space.

1. Introduction

Many problems in statistics, machine learning or signal processing can be formulated as high-dimensional convex optimization problems [3, 9, 41, 43, 47, 48]. These optimization problems typically involve a smooth term F and a nonsmooth regularization G, and are often solved using a (variant of) the proximal Stochastic Gradient Descent (SGD) [2]. However, in many cases, G is not proximable, *i.e.*, its proximity operator does not admit a closed form expression.

In particular, structured regularizations [9, 17] like the total variation regularization over a graph [7, 14, 21, 51] or the overlapping group lasso [3] are known to have an expensive proximity operator [45]. Another example is the case of affine constraints on the optimization problem. This corresponds to G being an indicator function and the proximity operator of G being the projection onto the constraints space. This projection requires the resolution of a high-dimensional linear system [4] often intractable. The context of decentralized optimization [16, 52], in which a network of computing agents aims at jointly minimizing an objective function by performing local computations and exchanging information along the edges, is a particular case of the context of linearly constrained optimization. In this particular case, projecting onto the constraints space is equivalent to averaging across the network, which is prohibited. Finally, when G is a sum of several regularizers, G is not

1. This paper is a short version of [46]

proximable even if the regularizers are proximable, because the proximity operator is not linear.

Although in these examples G is not proximable, G takes the form $G = R + H \circ L$ where R, H are proximable and L is a linear operator[2]. Therefore, we study the problem

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

where $\mathcal{X}$ is a real Hilbert space, F is a smooth convex function, R, H are nonsmooth convex functions and L is a linear operator.

Related works. *Splitting algorithms*: Algorithms allowing to minimize a function involving several nonsmooth proximable terms are called splitting algorithms. At the core of splitting algorithms is the Douglas-Rachford (or ADMM) algorithm [22, 36] which is, under reasonable assumptions, the only splitting algorithm that can minimize the sum of two nonsmooth functions $R + H$ [44]. To minimize $G = R + H \circ L$, the Douglas-Rachford algorithm can be generalized to the Primal Dual Hybrid Gradient (PDHG) algorithm, also called Chambolle-Pock algorithm [8]. Behind the success of PDHG is the ability to handle such a composite function G and hence the regularizations mentioned above. However, in signal processing and machine learning applications, the objective function usually involves a smooth data fitting term F. In order to cover these applications, splitting algorithms like Condat–Vũ [13, 50] and PD3O [53] were proposed to solve the Problem (1). These algorithms are primal dual in nature, *i.e.*, their iterates take the form $(x^k, y^k) \in \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y}$ is another real Hilbert space, x^k converges to a solution of Problem (1) and y^k converges to a solution of a dual of Problem (1).

Stochastic splitting algorithms: In machine learning applications, the gradient of F is often intractable and replaced by a cheaper stochastic gradient. These stochastic gradients can be classified in two classes: variance reduced (VR) stochastic gradients [19, 20, 23, 29] and generic stochastic gradients, see *e.g.* [33, 39]. VR stochastic gradients are stochastic gradient estimators of the full gradient that ensure convergence to an exact solution, as for deterministic algorithms. The variance reduction enables to speedup stochastic algorithms and eventually recover the convergence rates of their deterministic counterparts. In the case where $L = I$, Problem (1) was considered with generic stochastic gradients in [54] and with VR stochastic gradients in [42]. In the general case $L \neq I$ that is of interest in this paper, the resolution of (1) was considered with a generic stochastic gradient in [5, 55].

Contributions and technical challenges. In this paper we consider the resolution of Problem (1) with VR stochastic gradients, which enables for faster convergence compared to non VR approaches *e.g.* [55].

More precisely, we propose a new algorithm called Primal–Dual Davis–Yin (PDDY) to solve (1). This algorithm is obtained as a carefully designed instance of the DYS involving operators which are monotone under a metric depending on L. This DYS representation enables us to prove convergence rates for PDDY, a task which would be lengthy and technical if such a representation was not obtained prior to proving the convergence rates. We analyze PDDY with a deterministic gradient and with a variance reduced stochastic gradient. Both settings are cast into a single assumption which can be elegantly plugged into our analysis

2. In these contexts, $H \circ L$ is not proximable as well (the symbol $\circ$ stands for the composition of functions).

of PDDY, thanks to the flexibility of our framework. In the supplementary material, we also analyze PD3O [53] with a VR stochastic gradient. Our nonasymptotic results w.r.t PD3O and PDDY have recently been accelerated in the case of deterministic gradients [16]. Moreover, we show how the PD3O algorithm and the Condat–Vũ algorithms can also be seen as instances of DYS involving monotone operators. Such representation was not known for the Condat–Vũ algorithms.

One byproduct of our results is the discovery of one of the first linearly converging algorithm for the minimization of a smooth strongly convex function under affine constraints [37], without projecting onto the constraints space. In the particular case where a full gradient is used and L^*L is a gossip matrix [52], this algorithm leads to a new decentralized algorithm whose complexity competes with optimization algorithms specifically designed for the decentralized optimization problem, see Table 1 in [52]. Our decentralized algorithm has recently led to an optimal decentralized algorithm [32].

2. Primal–Dual Formulations and Optimality Conditions

The necessary notions and notations of convex analysis and operator theory are introduced in the Appendix. Let $\mathcal{X}$ and $\mathcal{Y}$ be finite-dimensional real Hilbert spaces, $L : \mathcal{X} \to \mathcal{Y}$ be a linear operator, $F, R \in \Gamma_0(\mathcal{X})$, and $H \in \Gamma_0(\mathcal{Y})$. We assume that F is ν-smooth, for some $\nu > 0$. We assume, as usual, that there exists $x^\star \in \mathcal{X}$ such that $0 \in \nabla F(x^\star) + \partial R(x^\star) + L^*\partial H(Lx^\star)$. Then $x^\star$ is solution to (1). Therefore, there exists $y^\star \in \partial H(Lx^\star) \subset \mathcal{Y}$ such that $0 \in M(x^\star, y^\star)$, where M is the set-valued operator defined by

$$M(x,y) \coloneqq \begin{bmatrix} \nabla F(x) + \partial R(x) & + L^*y \\ -Lx & + \partial H^*(y) \end{bmatrix}. \tag{2}$$

Conversely, for every solution to $0 \in M(x^\star, y^\star)$, $x^\star$ is a solution to (1) and $y^\star \in \arg\min(F + R)^* \circ (-L^*) + H^*$.

Finally, one can check that the operator M defined in (2) is monotone. Moreover, we have

$$M(x,y) = \underbrace{\begin{bmatrix} \partial R(x) \\ 0 \end{bmatrix}}_{\coloneqq A(x,y)} + \underbrace{\begin{bmatrix} L^*y \\ -Lx + \partial H^*(y) \end{bmatrix}}_{\coloneqq B(x,y)} + \underbrace{\begin{bmatrix} \nabla F(x) \\ 0 \end{bmatrix}}_{\coloneqq C(x,y)}, \tag{3}$$

and each term at the right hand side of (3) is maximal monotone, see Corollary 25.5 in [4].

3. The proposed PDDY Algorithm

We now set $\mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are the spaces defined in Section 2. Solving the optimization problem (1) boils down to finding a zero $(x^\star, y^\star)$ of the monotone operator M defined in (2). Since $M = A + B + C$, where the operator C is cocoercive, a natural idea is to apply the Davis–Yin splitting (DYS) algorithm [18], shown above. More precisely, if $\gamma < 2/\nu$, the iterates (z^k, u^k, v^k) of the algorithm DYS(A, B, C) or DYS(B, A, C) converge to some fixed point $(z^\star, u^\star, v^\star)$ such that $u^\star = z^\star$ and $z^\star$ is a zero of M.

Davis–Yin Algorithm DYS(A, B, C) [18]

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

Stochastic PDDY algorithm (proposed)
(deterministic version: $g^{k+1} = \nabla F(x^k)$)

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

However, the resolvent of B is often intractable. In this section, we show that preconditioning is the solution; that is, we exhibit a positive definite linear operator P, such that DYS$(P^{-1}A, P^{-1}B, P^{-1}C)$ and DYS$(P^{-1}B, P^{-1}A, P^{-1}C)$ are tractable.

Let $\gamma > 0$ and $\tau > 0$ be real parameters such that $\gamma\tau\|L\|^2 < 1$. Consider the positive definite operator

$$P := \begin{bmatrix} I & 0 \\ 0 & \frac{\gamma}{\tau} I - \gamma^2 L L^* \end{bmatrix}. \tag{4}$$

Since A, B, C are maximal monotone in $\mathcal{Z}$, $P^{-1}A, P^{-1}B, P^{-1}C$ are maximal monotone in $\mathcal{Z}_P$. Moreover, $P^{-1}C$ is $1/\nu$-cocoercive in $\mathcal{Z}_P$. Importantly, we have:

$$P^{-1}C : (x, y) \mapsto \big(\nabla F(x), 0\big), \quad J_{\gamma P^{-1}B} : (x, y) \mapsto \big(\operatorname{prox}_{\gamma R}(x), y\big), \tag{5}$$

$$J_{\gamma P^{-1}A} : (x, y) \mapsto (x', y'), \quad \text{where} \left\lfloor \begin{array}{l} y' = \operatorname{prox}_{\tau H^*}\big(y + \tau L(x - \gamma L^* y)\big) \\ x' = x - \gamma L^* y'. \end{array} \right. \tag{6}$$

If we plug these explicit steps into the Davis–Yin algorithm DYS$(P^{-1}A, P^{-1}B, P^{-1}C)$, we obtain the PD3O algorithm, see [53]. We propose to plug these explicit steps into the DYS$(P^{-1}B, P^{-1}A, P^{-1}C)$, and we identify the variables as $v^k = (p^k, q^k)$, $z^k = (x^k, y^k)$, $u^k = (s^k, d^k)$ and the fixed points as $v^\star = (p^\star, q^\star)$, $z^\star = (x^\star, y^\star)$, $u^\star = (s^\star, d^\star)$. Hence, $s^\star = x^\star$ is a solution of Problem (1) and $d^\star = y^\star$ is a solution to the dual problem. After some simplifications, we obtain the new Primal–Dual Davis–Yin (PDDY) algorithm, shown above, for Problem (1).

The convergence of PDDY is a consequence of the convergence of DYS$(P^{-1}B, P^{-1}A, P^{-1}C)$ along with the fact that the zeros of $P^{-1}M = P^{-1}A + P^{-1}B + P^{-1}C$ are the zeros of $M = A + B + C$.

Theorem 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 $(s^k)_{k\in\mathbb{N}}$ (resp. the sequence $(q^k)_{k\in\mathbb{N}}$) generated by the PDDY Algorithm converge to some solution $x^\star$ to Problem (1) (resp. some $y^\star \in \arg\min(F + R)^* \circ (-L^*) + H^*$).*

Note that PDDY converges for larger step sizes than the Condat–Vũ algorithms, see [13].

4. Non-Asymptotic Analysis of the stochastic PDDY algorithm

In the stochastic version of PDDY, the gradient $\nabla F(x^k)$ is replaced by a stochastic gradient g^{k+1}. More precisely, we consider a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_k)_k, \mathbb{P})$, an $(\mathscr{F}_k)$-adapted stochastic process $(g^k)_k$, we denote by $\mathbb{E}$ the mathematical 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 $(\mathscr{F}_k)_k$-adapted stochastic process denoted by $(\sigma_k)_k$, 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 a consequence of the smoothness of F and the choice of the stochastic gradient estimator, see [23]. Assumption 1 is satisfied by several stochastic gradient estimators used in machine learning, including the full gradient, some kinds of coordinate descent [25], variance reduction [20, 24, 27, 31], and also compressed gradients used to reduce the communication cost in distributed optimization [28], see Table 1 in [23].

We now analyze the proposed Stochastic PDDY Algorithm, shown above. We obtain sublinear convergence if M is not strongly monotone (Theorem 2) and linear convergence if M is strongly monotone (Theorem 3).

Theorem 2 (M monotone) *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,*

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

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 $\bar{s}^{k+1} = \frac{1}{k}\sum_{j=1}^{k} s^j$.

Theorem 3 (M strongly monotone) *Suppose that Assumption 1 holds. Also, suppose that H is $1/\mu_{H^*}$-smooth, F is μ_F-strongly convex and R is μ_R-strongly convex, where $\mu_R > 0$ and $\mu_{H^*} > 0$. For every $\kappa > \beta/\rho$ and every $\gamma, \tau > 0$ such that $\gamma \leq 1/(\alpha + \kappa\delta)$, $\gamma\tau\|L\|^2 < 1$ and $\gamma^2 \leq \frac{\mu_{H^*}}{\|L\|^2\mu_R}$, define $\eta := 2\left(\mu_{H^*} - \gamma^2\|L\|^2\mu_R\right) \geq 0$,*

$$V^k := (1+\gamma\mu_R)\|p^k - p^\star\|^2 + (1+\tau\eta)\|y^k - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\sigma_k^2, \tag{7}$$

and

$$r := \max\left(\frac{1}{1+\gamma\mu_R}, 1 - \rho + \frac{\beta}{\kappa}, \frac{1}{1+\tau\eta}\right) \tag{8}$$

Then,

$$\mathbb{E}V^k \leq r^k V^0. \tag{9}$$

Note that Theorem 3 does not assume R smooth, contrary to its analogue for PD3O, see [53] or Theorem 9 in the Appendix.

Now, we consider the particular case $R \equiv 0$ and $H : y \mapsto (0 \text{ if } y = b, +\infty \text{ else})$, for some $b \in \mathrm{ran}(L)$. In this case, Problem (1) boils down to $\min_x F(x)$ s.t. $Lx = b$. Moreover, this instance of the stochastic PDDY algorithm does not make use of projections onto the affine space $\{x \in \mathcal{X}, Lx = b\}$ (it only makes calls to L and L^*) while converging linearly.

Theorem 4 *Suppose that Assumption 1 holds, that F is μ_F-strongly convex, for some $\mu_F > 0$, and that $y^0 \in \mathrm{ran}(L)$. Let $y^\star$ be the unique element of $\mathrm{ran}(L)$ such that $\nabla F(x^\star) + L^* y^\star = 0$, and $\omega(L^*L) > 0$ be the smallest positive eigenvalue of L^*L. For every $\kappa > \beta/\rho$ and every $\gamma, \tau > 0$ such that $\gamma \leq 1/\alpha + \kappa\delta$ and $\gamma\tau\|L\|^2 < 1$, we define*

$$V^k := \|x^k - x^\star\|^2 + (1 + \tau\gamma\omega(L^*L))\,\|y^k - y^\star\|^2_{\gamma,\tau} + \kappa\gamma^2\mathbb{E}\sigma_k^2, \tag{10}$$

and

$$r := \max\left(1 - \gamma\mu_F, 1 - \rho + \frac{\beta}{\kappa}, \frac{1}{1 + \tau\gamma\omega(L^*L)}\right) < 1. \tag{11}$$

Then, for every $k \geq 0$,

$$\mathbb{E}V^k \leq r^k V^0. \tag{12}$$

Furthermore, this instance of the stochastic PDDY algorithm can be written using $W = L^*L$, $c = L^*b$ and primal variables in $\mathcal{X}$ only; this version, called PriLiCoSGD, is shown in the Appendix. Now, consider that $F = \frac{1}{M}\sum_{m=1}^M F_m$ is a finite sum of functions, that W is a gossip matrix of a network with M nodes [52], and that $c = 0$. In this case, PriLiCoSGD is a new decentralized algorithm. Theorem 4 applies and shows that, with the full gradient, ε-accuracy is reached after $\mathcal{O}(\max(\kappa, \chi)\log(1/\varepsilon))$ iterations, where κ is the condition number of F and $\chi = \|W\|/\omega(W)$. This complexity is better or equivalent to the one of recently proposed deterministic decentralized algorithms, like EXTRA, DIGing, NIDS, NEXT, Harness, Exact Diffusion, see Table 1 of [52], [35, Theorem 1] and [1]. With a stochastic gradient, the rate of our algorithm is also better than [38, Equation 99]. Finally, our decentralized algorithm has been accelerated in a recent paper to obtain the first optimal first-order algorithm for smooth and strongly convex decentralized optimization [32]. Their main complexity result is based on an extension of Theorem 4.

References

[1] S. A Alghunaim, E. K Ryu, K. Yuan, and A. H Sayed. Decentralized proximal gradient algorithms with linear convergence rates. *arXiv preprint arXiv:1909.06479*, 2019.

[2] Y. F Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *J. Mach. Learn. Res.*, 18(1):310–342, 2017.

[3] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.*, 4(1):1–106, 2012.

[4] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, New York, 2nd edition, 2017.

[5] P. Bianchi, W. Hachem, and A. Salim. A fully stochastic primal-dual algorithm. *Optimization Letters*, pages 1–10, 2020.

[6] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[7] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM J. Imaging Sci.*, 3(3):492–526, 2010.

[8] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *J. Math. Imaging Vision*, 40(1):120–145, May 2011.

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

[10] A. Chambolle and T. Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Math. Program.*, 159(1–2):253–287, September 2016.

[11] C.-C. Chang and C.-J. Lin. LibSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[12] P. L. Combettes, L. Condat, J.-C. Pesquet, and B. C. Vũ. A forward–backward view of some primal–dual optimization methods in image recovery. In *Proc. of IEEE ICIP*, Paris, France, October 2014.

[13] L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *J. Optim. Theory Appl.*, 158(2): 460–479, 2013.

[14] L. Condat. Discrete total variation: New definition and minimization. *SIAM J. Imaging Sci.*, 10(3):1258–1290, 2017.

[15] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi. Proximal splitting algorithms: A tour of recent advances, with new twists. *preprint arXiv:1912.00137*, 2019.

[16] L. Condat, G. Malinovsky, and P. Richtárik. Distributed proximal splitting algorithms with rates and acceleration. *preprint arXiv:2010.00952*, 2020.

[17] D. Cremers, T. Pock, K. Kolev, and A. Chambolle. Convex relaxation techniques for segmentation, stereo and multiview reconstruction. In *Markov Random Fields for Vision and Image Processing*. MIT Press, 2011.

[18] D. Davis and W. Yin. A three-operator splitting scheme and its optimization applications. *Set-Val. Var. Anal.*, 25:829–858, 2017.

[19] A. Defazio. A simple practical accelerated method for finite sums. In *Advances in Neural Information Processing Systems (NIPS)*, pages 676–684, 2016.

[20] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1646–1654, 2014.

[21] J. Duran, M. Moeller, C. Sbert, and D. Cremers. Collaborative total variation: A general framework for vectorial TV models. *SIAM J. Imaging Sci.*, 9(1):116–151, 2016.

[22] R. Glowinski and A. Marroco. 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, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975.

[23] E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of SGD: Variance reduction, sampling, quantization and coordinate descent. In *Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS)*, Palermo, Sicily, Italy, June 2020. to appear.

[24] R. M Gower, P. Richtárik, and F. Bach. Stochastic quasi-gradient methods: Variance reduction via Jacobian sketching. *Math. Program.*, 2020.

[25] F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2082–2093, 2018.

[26] B. S. He and X. M Yuan. Convergence analysis of primal–dual algorithms for a saddle-point problem: from contraction perspective. *SIAM J. Imaging Sci.*, 5:119–149, 2012.

[27] T. Hofmann, A. Lucchi, S. Lacoste-Julien, and B. McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2305–2313, 2015.

[28] S. Horváth, D. Kovalev, K. Mishchenko, S. Stich, and P. Richtárik. Stochastic distributed learning with gradient quantization and variance reduction. *preprint arXiv:1904.05115*, 2019.

[29] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 315–323, 2013.

[30] N. Komodakis and J.-C. Pesquet. Playing with duality: An overview of recent primal–dual approaches for solving large-scale optimization problems. *IEEE Signal Process. Mag.*, 32(6):31–54, November 2015.

[31] D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proc. of Int. Conf. Algo. Learn. Theory (ALT)*, 2020.

[32] D. Kovalev, A. Salim, and P. Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. to appear.

[33] G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning.* Springer, 2020.

[34] Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010.

[35] H. Li and Z. Lin. Revisiting extra for smooth distributed optimization. *arXiv preprint arXiv:2002.10110*, 2020.

[36] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964–979, 1979.

[37] K. Mishchenko and P. Richtárik. A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions. *arXiv preprint arXiv:1905.11535*, 2019.

[38] A. Mokhtari and A. Ribeiro. DSA: Decentralized double stochastic averaging gradient algorithm. *J. Mach. Learn. Res.*, 17(1):2165–2199, 2016.

[39] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459, 2011.

[40] D. O'Connor and L. Vandenberghe. On the equivalence of the primal-dual hybrid gradient method and Douglas–Rachford splitting. *Math. Program.*, 79:85–108, 2020.

[41] D. P. Palomar and Y. C. Eldar, editors. *Convex Optimization in Signal Processing and Communications*. Cambridge University Press, 2009.

[42] F. Pedregosa, K. Fatras, and M. Casotto. Proximal splitting meets variance reduction. In *Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS)*, pages 1–10, 2019.

[43] N. G. Polson, J. G. Scott, and B. T. Willard. Proximal algorithms in statistics and machine learning. *Statist. Sci.*, 30(4):559–581, 2015.

[44] E. K Ryu. Uniqueness of drs as the 2 operator resolvent-splitting and impossibility of 3 operator resolvent-splitting. *Math. Program.*, 182(1):233–273, 2020.

[45] A. Salim, P. Bianchi, and W. Hachem. Snake: a stochastic proximal gradient algorithm for regularized problems over large graphs. *IEEE Trans. Automat. Contr.*, 2019.

[46] A. Salim, L. Condat, K. Mishchenko, and P. Richtárik. Dualize, split, randomize: Fast nonsmooth optimization algorithms. *arXiv preprint arXiv:2004.02635*, 2020.

[47] J.-L. Starck, F. Murtagh, and J. Fadili. *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*. Cambridge University Press, 2010.

[48] G. Stathopoulos, H. Shukla, A. Szucs, Y. Pu, and C. N. Jones. Operator splitting methods in control. *Foundations and Trends in Systems and Control*, 3(3):249–362, 2016.

[49] J. K. Tay, J. Friedman, and R. Tibshirani. Principal component-guided sparse regression. *preprint arXiv:1810.04651*, 2018.

[50] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Adv. Comput. Math.*, 38(3):667–681, April 2013.

[51] Y.-X. Wang, J. Sharpnack, A. Smola, and R. Tibshirani. Trend filtering on graphs. *J. Mach. Learn. Res.*, 17:1–41, 2016.

[52] J. Xu, Y. Tian, Y. Sun, and G. Scutari. Distributed algorithms for composite optimization: Unified and tight convergence analysis. *arXiv preprint arXiv:2002.11534*, 2020.

[53] M. Yan. A new Primal–Dual algorithm for minimizing the sum of three functions with a linear operator. *J. Sci. Comput.*, 76(3):1698–1717, September 2018.

[54] A. Yurtsever, B. C. Vu, and V. Cevher. Stochastic three-composite convex minimization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4329–4337, 2016.

[55] R. Zhao and V. Cevher. Stochastic three-composite convex minimization with a linear operator. In *Proc. of Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2018.

Appendix

Contents

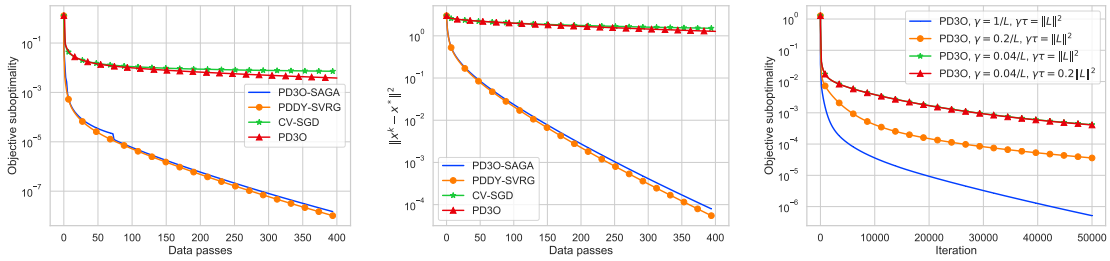

Figure 1: Results for the PCA-Lasso experiment. Left: convergence in the objective, middle: convergence in norm, right: the effect of the stepsizes.

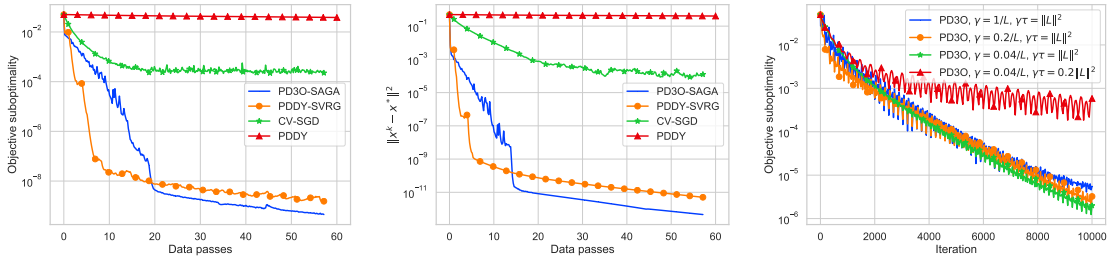

Figure 2: Results for the MNIST experiment. Left: convergence in the objective, middle: convergence in norm, right: the effect of the stepsizes.

Appendix A. Experiments

In this section, we present numerical experiments for the PDDY, PD3O and Condat–Vũ (CV) [13, Algorithm 3.1] algorithms. SGD was always used with a small γ, such as $\frac{0.01}{\nu}$. where ν is the smoothness constant of F. For stochastic methods, we used a batch size of 16 for better parallelism, while the sampling type is specified in the figures. The stepsizes were tuned with log-grid-search for all methods. We used closed-form expressions to compute ν for all problems and tuned the stepsizes for all methods by running logarithmic grid search with factor 1.5 over multiples of $\frac{1}{\nu}$.

We observed that the performances of these algorithms are nearly identical, when the same stepsizes are used, so we do not provide their direct comparison in the plots. Instead, we 1) compare different stochastic oracles, 2) illustrate how convergence differs in functional suboptimality and distances, and 3) show how the stepsizes affect the performance.

PCA-Lasso In a recent work [49, Eq. (12)] the following difficult PCA-based Lasso problem was introduced: $\min_x \frac{1}{2}\|Wx - a\|^2 + \lambda\|x\|_1 + \lambda_1 \sum_{i=1}^m \|L_i x\|$, where $W \in \mathbb{R}^{n\times p}$, $a \in \mathbb{R}^n$, $\lambda, \lambda_1 > 0$ are given. We generate 10 matrices L_i randomly with standard normal i.i.d. entries, each with 20 rows. W and y are taken from the 'mushrooms' dataset from the libSVM package [11]. We chose $\lambda = \frac{\nu}{10n}$ and $\lambda_1 = \frac{2\nu}{nm}$, where ν, the smoothness of F, is needed to compensate for the fact that we do not normalize the objective.

MNIST with Overlapping Group Lasso Now 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 $a \in \{0,1\}^n$, $F(x) = \frac{1}{n}\sum_{i=1}^n f_i(x) + \frac{\lambda}{2}\|x\|^2$ is a finite sum,

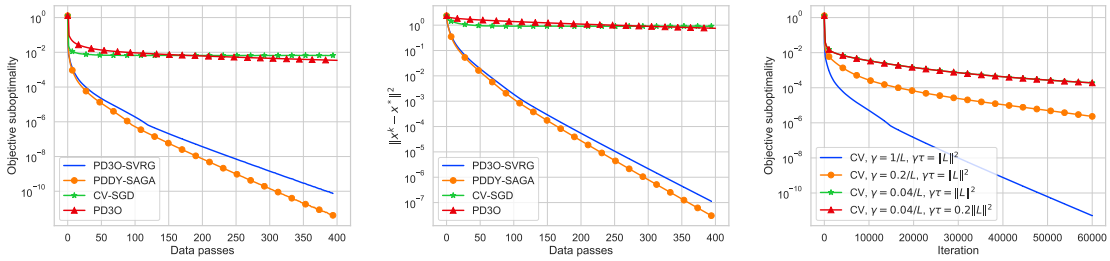

Figure 3: Results for the Fused Lasso experiment. Left: convergence with respect to the objective function, middle: convergence in norm, right: illustration of the effect of the stepsizes.

$f_i(x) = -\big(a_i \log\big(h(w_i^\top x)\big) + (1 - a_i) \log\big(1 - h(w_i^\top x)\big)\big)$, where, $\lambda = \frac{2\nu}{n}$, $w_i \in \mathbb{R}^p$ is the i-th row of W and $h : t \to 1/(1 + e^{-t})$ is the sigmoid function. The nonsmooth regularizer, in turn, 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 datasetw [34] 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.

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 smooth objective $F(x) = \frac{1}{2}\|Wx - a\|^2$. This function 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 use the 'mushrooms' dataset from the libSVM package. Our numerical findings for this problem are very similar to the ones for PCA-Lasso. In particular, larger values of γ seem to perform significantly better and the value of the objective function does not oscillate, unlike in the MNIST experiment. The results are shown in Figure 3. The proposed Stochastic PDDY algorithm with the SAGA estimator performs best in this setting.

Summary of results We can see from the plots that stochastic updates make the convergence extremely faster, sometimes even without variance reduction. The stepsize plots suggest that it is best to keep $\gamma\tau$ close to $\frac{1}{\|L\|^2}$, while the optimal value of γ might sometimes be smaller than $\frac{1}{\nu}$. This is especially clearly seen from the fact that SGD works sufficiently fast even despite using γ inversely proportional to the number of iterations.

Appendix B. Mathematical Background

We introduce some notions of convex analysis and operator theory, see the textbooks [4, 6] for more details. In the paper, all Hilbert spaces are supposed of finite dimension.

B.1. Convex functions

Let $\mathcal{Z}$ be a real Hilbert space, with its inner product $\langle \cdot\,, \cdot \rangle$ and norm $\|\cdot\| = \langle \cdot\,, \cdot \rangle^{1/2}$. Let $G : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$ be a convex function. The domain of G is the convex set $\operatorname{dom} G = \{z \in \mathcal{Z} \;:\; G(z) \neq +\infty\}$. G is proper if its domain is nonempty and lower semicontinuous if the convex set $\{z \in \mathcal{Z} \;:\; G(z) \leq \ell\}$ is closed, for every $\ell \in \mathbb{R}$. We denote by $\Gamma_0(\mathcal{Z})$ the set of convex, proper, lower semicontinuous functions from $\mathcal{Z}$ to $\mathbb{R} \cup \{+\infty\}$. We define the subdifferential of G as the set-valued operator $\partial G : z \in \mathcal{Z} \mapsto \{y \in \mathcal{Z} \;:\; (\forall z' \in \mathcal{Z})\ G(z) + \langle z' - z, y \rangle \leq G(z')\}$. If G is differentiable at $z \in \mathcal{Z}$, $\partial G(z) = \{\nabla G(z)\}$, where $\nabla G(z)$ denotes the gradient of G at z. In this case, the Bregman divergence of G is defined by

$$D_G(x, x') \coloneqq G(x) - G(x') - \langle \nabla G(x'), x - x' \rangle. \tag{13}$$

Moreover, G is ν-smooth if it is differentiable on $\mathcal{Z}$ and ∇G is ν-Lipschitz continuous, for some $\nu > 0$. We denote by G^* the conjugate of G, defined by $G^* : z \mapsto \sup_{z' \in \mathcal{Z}} \{\langle z, z' \rangle - G(z')\}$, which belongs to $\Gamma_0(\mathcal{Z})$. We define the proximity operator of G as the single-valued operator $\operatorname{prox}_G : z \in \mathcal{Z} \mapsto \arg\min_{z' \in \mathcal{Z}} \left\{G(z') + \frac{1}{2}\|z - z'\|^2\right\}$. Finally, given any $b \in \mathcal{Z}$, we define the indicator function $\iota_b : z \mapsto \{0 \text{ if } z = b,\ +\infty \text{ else}\}$, which belongs to $\Gamma_0(\mathcal{Z})$.

B.2. Monotone operators

Consider a set-valued operator $M : \mathcal{Z} \rightrightarrows \mathcal{Z}$. 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 $\operatorname{zer}(M) = M^{-1}(0) = \{z \in \mathcal{Z}, 0 \in M(z)\}$. The operator M is monotone if $\langle w - w', z - z' \rangle \geq 0$, whenever $u \in A(z)$ and $u' \in A(z')$, and strongly monotone if there exists $\mu > 0$, such that $\langle w - w', z - z' \rangle \geq \mu \|z - z'\|^2$. The resolvent operator of M is defined by $J_M = (I + M)^{-1}$, where I denotes the identity. If M is monotone, then $J_M(z)$ is either empty or single-valued. M is maximal monotone if $J_M(z)$ is single-valued, for every $z \in \mathcal{Z}$. We identify single-valued operators as operators from $\mathcal{Z}$ to $\mathcal{Z}$. If $G \in \Gamma_0(\mathcal{Z})$, then ∂G is maximal monotone, $J_{\partial G} = \operatorname{prox}_G$, $\operatorname{zer}(\partial G) = \arg\min G$ and $(\partial G)^{-1} = \partial G^*$.

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

Let $\mathcal{X}, \mathcal{Y}$ be real Hilbert spaces and let $L : \mathcal{X} \to \mathcal{Y}$ be a linear operator. The adjoint of L is denoted by $L^* : \mathcal{Y} \to \mathcal{X}$, and the operator norm of L is $\|L\| = \sup\{\|Lx\|, x \in \mathcal{X}, \|x\| \leq 1\}$. The largest eigenvalue of LL^* is $\|LL^*\| = \|L\|^2 = \|L^*\|^2$. Let $P : \mathcal{Z} \to \mathcal{Z}$ be a linear and symmetric operator ($P^* = P$). P is positive semidefinite if $\langle Pz, z \rangle \geq 0$, for every $z \in \mathcal{Z}$, and positive definite 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 is defined by $\|z\|_P^2 = \langle z, z \rangle_P$. We denote by $\mathcal{Z}_P$ the space $\mathcal{Z}$ endowed with $\langle \cdot, \cdot \rangle_P$. Finally, we denote $\|\cdot\|_{\gamma,\tau}$ the norm induced by $\frac{\gamma}{\tau} I - \gamma^2 LL^*$ on $\mathcal{Y}$.

B.3. Primal–Dual Optimality

Let $x^\star$ be a minimizer of Problem (1). Assuming a standard qualification condition, for instance that 0 belongs to the relative interior of $\operatorname{dom}(H) - L \operatorname{dom}(R)$, then for every

$x \in \mathcal{X}$,

$$\partial(F + R + H \circ L)(x) = \nabla F(x) + \partial R(x) + L^* \partial H(Lx),$$

see for instance Theorem 16.47 of [4]. Then,

$$\begin{aligned}
& x^\star \in \arg\min_{x \in \mathcal{X}} \{F(x) + R(x) + H(Lx)\} \\
\Leftrightarrow\ & 0 \in \nabla F(x^\star) + \partial R(x^\star) + L^* \partial H(Lx^\star) \\
\Leftrightarrow\ & \exists y^\star \in \partial H(Lx^\star) \text{ such that } 0 \in \nabla F(x^\star) + \partial R(x^\star) + L^* y^\star \\
\Leftrightarrow\ & \exists y^\star \in \mathcal{Y} \text{ such that } 0 \in \nabla F(x^\star) + \partial R(x^\star) + L^* y^\star \text{ and } 0 \in -Lx^\star + \partial H^*(y^\star),
\end{aligned}$$

where we used $\partial H^* = (\partial H)^{-1}$. Moreover, such $y^\star \in \mathcal{Y}$ satisfies

$$0 \in -L\partial(F + R)^*(-L^* y^\star) + Lx^\star \text{ and } 0 \in -Lx^\star + \partial H^*(y^\star),$$

therefore $0 \in \partial(F + R)^* \circ (-L^*)(y^\star) + \partial H^*(y^\star)$ and

$$y^\star \in \arg\min (F + R)^* \circ (-L^*) + H^*.$$

In summary, there exist $r^\star \in \partial R(x^\star)$ and $h^\star \in \partial H^*(y^\star)$ such that

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \nabla F(x^\star) + r^\star & + L^* y^\star \\ -Lx^\star & + h^\star \end{bmatrix}. \tag{14}$$

In the sequel, we let $(x^\star, y^\star) \in \operatorname{zer}(M)$ and $r^\star, h^\star$ be any elements such that Equation (14) holds.

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

The inclusion (14) characterizes the first-order optimality conditions associated with the convex–concave Lagrangian function defined as

$$\mathscr{L}(x, y) \coloneqq (F + R)(x) - H^*(y) + \langle Lx, y \rangle. \tag{15}$$

For every $x \in \mathcal{X}$, $y \in \mathcal{Y}$, we define the *duality gap* at (x, y) as $\mathscr{L}(x, y^\star) - \mathscr{L}(x^\star, y)$. Then

Lemma 5 (Duality gap) *For every $x \in \mathcal{X}$, $y \in \mathcal{Y}$, we have*

$$\mathscr{L}(x, y^\star) - \mathscr{L}(x^\star, y) = D_F(x, x^\star) + D_R(x, x^\star) + D_{H^*}(y, y^\star). \tag{16}$$

Proof Using the optimality conditions (3), 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

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

Stochastic PD3O algorithm (proposed)
(deterministic version: $g^{k+1} = \nabla F(x^k)$)

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

PriLiCoSGD (proposed)
(deterministic version: $g^{k+1} = \nabla F(x^k)$)

Input: $x^0 \in \mathcal{X}, \gamma > 0, \tau > 0$
for $k = 0, 1, 2, \ldots$ **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

Summing the two last equations, we have

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

■

For every $x \in \mathcal{X}$, $y \in \mathcal{Y}$, Lemma 5 and the convexity of F, R, H^* imply that

$$\mathscr{L}(x^\star, y) \leq \mathscr{L}(x^\star, y^\star) \leq \mathscr{L}(x, y^\star). \tag{17}$$

So, the duality 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 $\min_{y\in\mathcal{Y}}(F + R)^*(-L^*y) + H^*(y)$, see Section 15.3 of [4]. The converse is true under mild assumptions, for instance strict convexity of the functions around $x^\star$ and $y^\star$.

Appendix C. Primal Dual Algorithms and their DYS representation

In this section, we show how other primal dual algorithms can be obtained as preconditioned instances of DYS, similarly to PDDY.

C.1. The PD3O Algorithm

As mentioned in Section 3, if we apply DYS($P^{-1}A, P^{-1}B, P^{-1}C$), then we recover exactly the PD3O algorithm proposed in [53]. 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 1.

We can note that in a recent work [40], the PD3O algorithm has been shown to be an instance of the Davis–Yin algorithm, with a different reformulation, which does not involve duality. Whether this connection could yield different insights on the PD3O algorithm is left for future investigation.

C.1.1. Non-Asymptotic Analysis of the stochastic PD3O algorithm

Similarly to the stochastic PDDY algorithm, we can obtain convergence rates for the stochastic PD3O with a VR stochastic gradient thanks to its DYS representation.

We obtain sublinear convergence if M is not strongly monotone (Theorem 6) and linear convergence if M is strongly monotone (Theorem 9).

Theorem 6 *Suppose that Assumption 1 holds. Let $\kappa := \beta/\rho$, $\gamma, \tau > 0$ be such that $\gamma \leq 1/2(\alpha + \kappa\delta)$ and $\gamma\tau\|L\|^2 < 1$. Set $V^0 := \|v^0 - v^\star\|_P^2 + \gamma^2\kappa\sigma_0^2$, where $v^0 = (p^0, y^0)$. Then,*

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

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

In the deterministic case $g^{k+1} = \nabla F(x^k)$, we recover the same rate as in [53, Theorem 2].

Remark 7 (Primal–Dual gap) *Deriving a similar bound on the stronger primal–dual gap $(F + R + H \circ L)(\bar{x}^k) + ((F + R)^* \circ -L + H^*)(\bar{y}^k)$ requires additional assumptions; for instance, even for the Chambolle–Pock algorithm, which is the particular case of the PD3O, PPDY and Condat–Vũ algorithm when $F = 0$, the best available result [10, Theorem 1] is not stronger than Theorem 6.*

Remark 8 (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 6 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, where the stochastic gradient satisfies Assumption 1. This covers coordinate descent and variance-reduced versions, as discussed previously.*

Theorem 9 (M strongly monotone and R smooth) *Suppose that Assumption 1 holds. Also, suppose that H is $1/\mu_{H^*}$-smooth, F is μ_F-strongly convex, and R is μ_R-strongly convex and λ-smooth, where $\mu := \mu_F + 2\mu_R > 0$ and $\mu_{H^*} > 0$. For every $\kappa > \beta/\rho$ and every $\gamma, \tau > 0$ such that $\gamma \leq 1/(\alpha + \kappa\delta)$ and $\gamma\tau\|L\|^2 < 1$, define*

$$V^k := \|p^k - p^\star\|^2 + (1 + 2\tau\mu_{H^*})\,\|y^k - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\sigma_k^2, \tag{18}$$

and

$$r := \max\left(1 - \frac{\gamma\mu}{(1+\gamma\lambda)^2}, \left(1 - \rho + \frac{\beta}{\kappa}\right), \frac{1}{1 + 2\tau\mu_{H^*}}\right). \tag{19}$$

Then,

$$\mathbb{E}V^k \leq r^k V^0. \tag{20}$$

In the deterministic case $g^{k+1} = \nabla F(x^k)$, we recover the same rate as in [53, Theorem 2], under the similar assumptions[3].

Since $\|x^k - x^\star\| \leq \|p^k - p^\star\|$, Theorem 9 also implies linear convergence of the primal variable x^k to $x^\star$, with same convergence rate.

3. Additionally, we correct some typos in the rate of [53].

Remark 10 (Particular case) *In the case where $R = H = 0$ and $L = 0$, then the Stochastic PD3O Algorithm boils down to Stochastic Gradient Descent (SGD), where the stochastic gradient oracle satisfies Assumption 1. Moreover, the value of r boils down to $r = \max\left(1-\gamma\mu, \left(1-\rho+\frac{\beta}{\kappa}\right)\right)$. Consider the applications of SGD covered by Assumption 1, and mentioned in Section 4. Then, as proved in [23], the value $r = \max\left(1-\gamma\mu, \left(1-\rho+\frac{\beta}{\kappa}\right)\right)$ matches the best known convergence rates for these applications, with an exception for some coordinate descent algorithms. However, if $H = 0$ and $L = 0$ but $R \neq 0$, then the Stochastic PD3O Algorithm boils down to Proximal SGD, and r boils down to $r = \max\left(1-\frac{\gamma\mu}{(1+\gamma\lambda)^2}, \left(1-\rho+\frac{\beta}{\kappa}\right)\right)$, whereas the best known rates for Proximal SGD under Assumption 1 is $\max\left(1-\gamma\mu, \left(1-\rho+\frac{\beta}{\kappa}\right)\right)$.*

C.2. The Condat–Vũ Algorithms

Let $\gamma > 0$ and $\tau > 0$ be real parameters. We define the operators

$$\bar{A}(x,y) = \begin{bmatrix} \partial R(x) + L^* y \\ -Lx \end{bmatrix},\ \bar{B}(x,y) = \begin{bmatrix} 0 \\ \partial H^*(y) \end{bmatrix},\ C(x,y) = \begin{bmatrix} \nabla F(x) \\ 0 \end{bmatrix},\ Q = \begin{bmatrix} K & 0 \\ 0 & I \end{bmatrix}, \tag{21}$$

where $K \coloneqq \frac{\gamma}{\tau} I - \gamma^2 L^* L$. Then, $M = \bar{A} + \bar{B} + C$. If $\gamma\tau\|L\|^2 < 1$, K and Q are positive definite. In that case, since $\bar{A}$, $\bar{B}$, C are maximal monotone in $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $Q^{-1}\bar{A}$, $Q^{-1}\bar{B}$, $Q^{-1}C$ are maximal monotone in $\mathcal{Z}_Q$. Moreover, we have:

$$Q^{-1}C : (x,y) \mapsto \big(K^{-1}\nabla F(x), 0\big), \quad J_{\gamma Q^{-1}\bar{B}} : (x,y) \mapsto \big(x, \operatorname{prox}_{\gamma H^*}(y)\big), \tag{22}$$

$$J_{\gamma Q^{-1}\bar{A}} : (x,y) \mapsto (x',y'), \quad \text{where} \left| \begin{array}{l} x' = \operatorname{prox}_{\tau R}\big((I - \tau\gamma L^* L)x - \tau L^* y\big) \\ y' = y + \gamma L x'. \end{array} \right. \tag{23}$$

We claim without proof that if we plug these explicit steps into the Davis–Yin algorithm $\mathrm{DYS}(Q^{-1}\bar{A}, Q^{-1}\bar{B}, Q^{-1}C)$ or $\mathrm{DYS}(Q^{-1}\bar{B}, Q^{-1}\bar{A}, Q^{-1}C)$, we recover the two forms of the Condat–Vũ algorithm [13, 50]; that is, Algorithms 3.1 and 3.2 of [13], respectively. The Condat–Vũ algorithm has the form of a primal–dual forward–backward algorithm [12, 15, 26, 30]. But we have just seen that it can be viewed as a primal–dual Davis–Yin algorithm, with a different metric, as well.

Appendix D. Proofs

D.1. Fundamental equality of the DYS Algorithm

The proofs of our non-asymptotic rates are a combination of the DYS representation of our algorithms along with the following general inequality w.r.t. the DYS algorithm.

Lemma 11 *Let $(v^k, z^k, u^k) \in \mathcal{Z}^3$ be the iterates of the DYS(A,B,C) algorithm, and $(v^\star, z^\star, u^\star) \in \mathcal{Z}^3$ be a fixed point of the DYS(A,B,C) algorithm:*

$$z^\star = J_{\gamma B}(v^\star), \quad u^\star = J_{\gamma A}\big(2z^\star - v^\star - \gamma C(z^\star)\big), \quad u^\star = z^\star. \tag{24}$$

Then, for every $k \geq 0$, there exist $b^k \in B(z^k), b^\star \in B(z^\star), a^{k+1} \in A(u^{k+1})$ and $a^\star \in A(u^\star)$ such that

$$\begin{aligned} \|v^{k+1} - v^\star\|^2 = \|v^k - v^\star\|^2 - 2\gamma\langle b^k - b^\star, z^k - z^\star\rangle - 2\gamma\langle C(z^k) - C(z^\star), z^k - z^\star\rangle \\ - 2\gamma\langle a^{k+1} - a^\star, u^{k+1} - u^\star\rangle - \gamma^2\|a^{k+1} + b^k - (a^\star + b^\star)\|^2 + \gamma^2\|C(z^k) - C(z^\star)\|^2. \end{aligned} \tag{25}$$

Proof Since $z^k = J_{\gamma B}(v^k)$, $z^k \in v^k - \gamma B(z^k)$ by definition of the resolvent operator. Therefore, there exists $b^k \in B(z^k)$ such that $z^k = v^k - \gamma b^k$. Similarly,

$$u^{k+1} \in 2z^k - v^k - \gamma C(z^k) - \gamma A(u^{k+1}) = v^k - 2\gamma b^k - \gamma C(z^k) - \gamma A(u^{k+1}).$$

Therefore, there exists $a^{k+1} \in A(u^{k+1})$ such that

$$\left\{ \begin{array}{l} z^k = v^k - \gamma b^k \\ u^{k+1} = v^k - 2\gamma b^k - \gamma C(z^k) - \gamma a^{k+1} \\ v^{k+1} = v^k + u^{k+1} - z^k. \end{array} \right. \tag{26}$$

Moreover,

$$v^{k+1} = v^k - \gamma b^k - \gamma C(z^k) - \gamma a^{k+1}. \tag{27}$$

Similarly, there exist $a^\star \in A(u^\star), b^\star \in B(z^\star)$ such that

$$\left\{ \begin{array}{l} z^\star = v^\star - \gamma b^\star \\ u^\star = v^\star - 2\gamma b^\star - \gamma C(z^\star) - \gamma a^\star \\ v^\star = v^\star + u^\star - z^\star, \end{array} \right. \tag{28}$$

and

$$v^\star = v^\star - \gamma b^\star - \gamma C(z^\star) - \gamma a^\star. \tag{29}$$

Therefore, using (27) and (29),

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

By expanding the last square at the right hand side, and by using (26) and (28) in the inner product, we get

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

Then, the last five terms at the right hand side simplify to

$$\gamma^2 \|C(z^k) - C(z^\star)\|^2 - \gamma^2 \|a^{k+1} + b^k - (a^\star + b^\star)\|^2,$$

and we get the result. ■

D.2. Proofs related to the Stochastic PDDY Algorithm

We start by proving Equation (6) using the notations of Section 3.

Lemma 12 *$J_{\gamma P^{-1}A}$ maps (x,y) to (x',y'), such that*

$$\left\lfloor \begin{array}{l} y' = \operatorname{prox}_{\tau H^*}\big(y + \tau L(x - \gamma L^* y)\big) \\ x' = x - \gamma L^* y'. \end{array} \right. \tag{30}$$

Proof Let (x,y) and $(x',y') \in \mathcal{Z}$, such that

$$P \begin{bmatrix} x'-x \\ y'-y \end{bmatrix} \in -\gamma \begin{bmatrix} & + L^* y' \\ -Lx' & + \partial H^*(y') \end{bmatrix},$$

where

$$P = \begin{bmatrix} I & 0 \\ 0 & \frac{\gamma}{\tau} I - \gamma^2 L L^* \end{bmatrix}.$$

We shall express (x',y') as a function of (x,y). First,

$$x' = x - \gamma L^* y'.$$

Moreover, y' is given by

$$\begin{aligned} \left(\frac{\gamma}{\tau} I - \gamma^2 L L^*\right)(y') &\in \left(\frac{\gamma}{\tau} I - \gamma^2 L L^*\right)(y) + \gamma L x' - \gamma \partial H^*(y') \\ &\in \left(\frac{\gamma}{\tau} I - \gamma^2 L L^*\right)(y) + \gamma L\left(x - \gamma L^* y'\right) - \gamma \partial H^*(y'). \end{aligned}$$

Therefore, the term $\gamma^2 L L^* y'$ disappears from both sides and

$$y' \in y - \gamma\tau L L^* y - \tau \partial H^*(y') + \tau L x.$$

Finally,

$$y - \gamma\tau L L^* y + \tau L x \in y' + \tau \partial H^*(y'),$$

and

$$y' = \operatorname{prox}_{\tau H^*}(y - \gamma\tau L L^* y + \tau L x).$$

■

Recall that the PDDY algorithm is equivalent to $\mathrm{DYS}(P^{-1}B, P^{-1}A, P^{-1}C)$. We denote by $v^k = (p^k, q^k)$, $z^k = (x^k, y^k)$, $u^k = (s^k, d^k)$ the iterates of $\mathrm{DYS}(P^{-1}B, P^{-1}A, P^{-1}C)$, where $p^k, x^k, s^k \in \mathcal{X}$ and $q^k, y^k, d^k \in \mathcal{Y}$.

Using (6), the step

$$z^k = J_{\gamma P^{-1}A}(v^k),$$

is equivalent to

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

Then, 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

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

Finally, the step

$$v^{k+1} = v^k + u^{k+1} - z^k$$

is equivalent to

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

Similarly, the fixed points $v^\star = (p^\star, q^\star), z^\star = (x^\star, y^\star), u^\star = (s^\star, d^\star)$ of $\mathrm{DYS}(P^{-1}B, P^{-1}A, P^{-1}C)$ satisfy

$$\left\{ \begin{array}{l} x^\star = p^\star - \gamma L^* y^\star \\ y^\star = \mathrm{prox}_{\tau H^*}\left((I - \tau\gamma LL^*)q^\star + \tau L p^\star\right) \\ s^\star = \mathrm{prox}_{\gamma R}\left(2x^\star - p^\star - \gamma \nabla F(x^\star)\right) \\ d^\star = 2y^\star - q^\star \\ p^\star = p^\star + s^\star - x^\star \\ q^\star = q^\star + d^\star - y^\star, \end{array} \right.$$

and the iterates of the stochastic PDDY algorithm satisfy

$$\left\lfloor \begin{array}{l} x^k = p^k - \gamma L^* y^k \\ y^k = \mathrm{prox}_{\tau H^*}\left((I - \tau\gamma LL^*)q^k + \tau L p^k\right) \\ s^{k+1} = \mathrm{prox}_{\gamma R}\left(2x^k - p^k - \gamma g^{k+1}\right) \\ d^{k+1} = 2y^k - q^k \\ p^{k+1} = p^k + s^{k+1} - x^k \\ q^{k+1} = q^k + d^{k+1} - y^k. \end{array} \right.$$

Lemma 13 *Suppose that (g^k) 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 11 for DYS($P^{-1}B, P^{-1}A, P^{-1}C$) using the norm induced by P, we have

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

Using

$$\begin{aligned}
A(z^k) &= \begin{bmatrix} L^* y^k \\ -Lx^k + \partial H^*(y^k) \end{bmatrix} \\
B(u^{k+1}) &= \begin{bmatrix} \partial R(s^{k+1}) \\ 0 \end{bmatrix} \\
C(z^k) &= \begin{bmatrix} g^{k+1} \\ 0 \end{bmatrix},
\end{aligned}$$

and

$$\begin{aligned}
A(z^\star) &= \begin{bmatrix} L^* y^\star \\ -Lx^\star + \partial H^*(y^\star) \end{bmatrix} \\
B(u^\star) &= \begin{bmatrix} \partial R(s^\star) \\ 0 \end{bmatrix} \\
C(z^\star) &= \begin{bmatrix} \nabla F(x^\star) \\ 0 \end{bmatrix},
\end{aligned}$$

we have,

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

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

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

Using the convexity of F,

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

Using Assumption 1,

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

■

D.2.1. Proof of Theorem 2

Using Lemma 13 and the convexity of F, R, H^*,

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

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

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

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,

$$\gamma \sum_{j=0}^{k-1} \mathbb{E}\left(D_F(x^k, x^\star) + D_{H^*}(y^k, y^\star) + D_R(s^{k+1}, s^\star)\right) \leq \mathbb{E}V^0. \tag{31}$$

We conclude using the convexity of the Bregman divergence in its first variable.

D.2.2. Proof of Theorem 3

We first use Lemma 13 along with the strong convexity of R, H^*. Note that $y^k = q^{k+1}$. We have

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

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 $-\|a+b\|^2 \leq -\frac{1}{2}\|a\|^2 + \|b\|^2$, we have

$$-\mathbb{E}_k\|s^{k+1} - s^\star\|^2 \leq -\frac{1}{2}\mathbb{E}_k\|p^{k+1} - p^\star\|^2 + \gamma^2\|L\|^2\mathbb{E}_k\|q^{k+1} - q^\star\|^2.$$

Hence,

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

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

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

Set

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

and

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

Then

$$\mathbb{E}_k V^{k+1} \leq rV^k.$$

D.3. Proofs related to the Stochastic PD3O Algorithm

Recall that the PD3O algorithm is equivalent to DYS($P^{-1}A, P^{-1}B, P^{-1}C$). We denote by $v^k = (p^k, q^k)$, $z^k = (x^k, y^k)$, $u^k = (s^k, d^k)$ the variables in DYS($P^{-1}A, P^{-1}B, P^{-1}C$), with $p^k, x^k, s^k \in \mathcal{X}$ and $q^k, y^k, d^k \in \mathcal{Y}$.

Then, the step

$$z^k = J_{\gamma P^{-1}B}(v^k),$$

is equivalent to

$$\left\lfloor \begin{array}{l} x^k = \mathrm{prox}_{\gamma R}(p^k) \\ y^k = q^k. \end{array} \right.$$

Using (6), the step

$$u^{k+1} = J_{\gamma P^{-1}A}(2z^k - v^k - \gamma P^{-1}C(z^k)),$$

is equivalent to

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

Finally, the step

$$v^{k+1} = v^k + u^{k+1} - z^k,$$

is equivalent to

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

Similarly, the fixed points $v^\star = (p^\star, q^\star), z^\star = (x^\star, y^\star), u^\star = (s^\star, d^\star)$ of DYS($P^{-1}A, P^{-1}B, P^{-1}C$) satisfy

$$\left\lfloor \begin{array}{l} x^\star = \mathrm{prox}_{\gamma R}(p^\star) \\ y^\star = q^\star \\ s^\star = (2x^\star - p^\star - \gamma \nabla F(x^\star)) - \gamma L^* d^\star \\ d^\star = \mathrm{prox}_{\tau H^*}\left((I - \gamma\tau LL^*)(2y^\star - q^\star) + \tau L(2x^\star - p^\star - \nabla F(x^\star))\right) \\ p^\star = p^\star + s^\star - x^\star \\ q^\star = q^\star + d^\star - y^\star. \end{array} \right.$$

and the iterates of the stochastic PD3O algorithm satisfy

$$\left\lfloor \begin{array}{l} x^k = \mathrm{prox}_{\gamma R}(p^k) \\ y^k = q^k \\ s^{k+1} = (2x^k - p^k - \gamma g^{k+1}) - \gamma L^* d^{k+1} \\ d^{k+1} = \mathrm{prox}_{\tau H^*}\left((I - \gamma\tau LL^*)(2y^k - q^k) + \tau L(2x^k - p^k - g^{k+1})\right) \\ p^{k+1} = p^k + s^{k+1} - x^k \\ q^{k+1} = q^k + d^{k+1} - y^k. \end{array} \right.$$

Lemma 14 *Assume that F is μ_F-strongly convex, for some $\mu_F \geq 0$, and that (g^k) satisfies Assumption 1. Then, the iterates of the Stochastic PD3O Algorithm satisfy*

$$\begin{aligned}\mathbb{E}_k\|v^{k+1} - v^\star\|_P^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 \leq \|v^k - v^\star\|_P^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2 \\ - 2\gamma(1 - \gamma(\alpha + \kappa\delta))D_F(x^k, x^\star) - \gamma\mu_F\|x^k - x^\star\|^2 \\ - 2\gamma\langle \partial R(x^k) - \partial R(x^\star), x^k - x^\star\rangle \\ - 2\gamma\mathbb{E}_k\langle \partial H^*(d^{k+1}) - \partial H^*(d^\star), d^{k+1} - d^\star\rangle \\ - \gamma^2\mathbb{E}_k\|P^{-1}A(u^{k+1}) + P^{-1}B(z^k) \\ - \left(P^{-1}A(u^\star) + P^{-1}B(z^\star)\right)\|_P^2.\end{aligned} \tag{32}$$

Proof Applying Lemma 11 for DYS$(P^{-1}A, P^{-1}B, P^{-1}C)$ using the norm induced by P we have

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

Using

$$\begin{aligned}A(u^{k+1}) &= \begin{bmatrix} L^*d^{k+1} \\ -Ls^{k+1} + \partial H^*(d^{k+1})\end{bmatrix} \\ B(z^k) &= \begin{bmatrix}\partial R(x^k) \\ 0\end{bmatrix} \\ C(z^k) &= \begin{bmatrix} g^{k+1} \\ 0\end{bmatrix},\end{aligned}$$

and

$$A(u^\star) = \begin{bmatrix} L^* d^\star \\ -Ls^\star + \partial H^*(d^\star) \end{bmatrix}$$
$$B(z^\star) = \begin{bmatrix} \partial R(x^\star) \\ 0 \end{bmatrix}$$
$$C(z^\star) = \begin{bmatrix} \nabla F(x^\star) \\ 0 \end{bmatrix},$$

we have

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

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,

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

■

D.3.1. Proof of Theorem 6

Using Lemma 14, convexity of F, R, H^*, and Lemma 5,

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

Recall that $1-\rho+\beta/\kappa=1$, $\gamma\le 1/2(\alpha+\kappa\delta)$. Set

$$V^k=\|v^k-v^\star\|_P^2+\kappa\gamma^2\sigma_k^2.$$

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,

$$\gamma\sum_{j=0}^{k-1}\mathbb{E}\left(\mathscr{L}(x^j,d^\star)-\mathscr{L}(x^\star,d^{j+1})\right)\le\mathbb{E}V^0.$$

We conclude using the convex-concavity of L.

D.3.2. Proof of Theorem 9

We first use Lemma 14 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

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

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

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

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 11 with $A = 0$, $C = 0$ and $B = \nabla R$, we obtain that if $x^k = \mathrm{prox}_{\gamma R}(p^k)$ and $x^\star = \mathrm{prox}_{\gamma R}(p^\star)$, 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,

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

Thus, set

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

and

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

Then,

$$\mathbb{E}_k V^{k+1} \leq rV^k.$$

D.4. Proof of Theorem 4

We first derive the following lemma:

Lemma 15 *Let $x \in \mathrm{ran}(L^*)$, the range space of L^*. There exists an unique $y \in \mathrm{ran}(L)$ such that $L^*y = x$. Moreover, for every $y \in \mathrm{ran}(L)$,*

$$\omega(L)\|y\|^2 \leq \|L^*y\|^2, \tag{33}$$

where $\omega(L)$ is the smallest positive eigenvalue of LL^* (or L^*L).

Proof Using basic linear algebra, $LL^*x = 0$ implies $L^*x \in \mathrm{ran}(L^*) \cap \ker(L)$ therefore $L^*x = 0$. Hence, $\ker(LL^*) \subset \ker(L^*)$ and therefore $\mathrm{ran}(L) \subset \mathrm{ran}(LL^*)$. Since LL^* is real symmetric, for every $y \in \mathrm{ran}(LL^*)$, $\langle y, LL^*y\rangle \geq \omega(L)\|y\|^2$, where $\omega(L)$ is the smallest positive eigenvalue of LL^*. Therefore, for every $y \in \mathrm{ran}(L)$, $\|L^*y\|^2 \geq \omega(L)\|y\|^2$. Moreover, $L^*y = 0$ implies $y = 0$ on $\mathrm{ran}(L)$, therefore there is at most one solution y in $\mathrm{ran}(L)$ to the equation $L^*y = x$. The existence of a solution follows from $x \in \mathrm{ran}(L^*)$. ■

Now, we prove Theorem 4. First, we define $y^\star$. In the case $R = 0$ and $H = \iota_b$, Equation (14) states that $\nabla F(x^\star) \in \mathrm{ran}(L^*)$. Using Lemma 15, there exists an unique

$y^\star \in \operatorname{ran}(L)$ such that $\nabla F(x^\star) + L^* y^\star = 0$. Noting that $y^\star = d^\star = q^\star$ and applying Lemma 14 with $\gamma \leq (\alpha + \kappa\delta)$,

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

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

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

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, using Lemma 15, $\omega(L)\|d^{k+1} - d^\star\|^2 \leq \|L^*d^{k+1} - L^*d^\star\|^2$. Since $q^{k+1} = d^{k+1} = y^{k+1}$ and $x^k = p^k$,

$$\begin{aligned}&\mathbb{E}_k\|x^{k+1} - x^\star\|^2 + (1 + \gamma\tau\omega(L))\mathbb{E}_k\|y^{k+1} - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\mathbb{E}_k\sigma_{k+1}^2 \\ &\leq (1 - \gamma\mu_F)\|x^k - x^\star\|^2 + \|y^k - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\left(1 - \rho + \frac{\beta}{\kappa}\right)\sigma_k^2.\end{aligned}$$

Setting

$$V^k = \|x^k - x^\star\|^2 + \big(1 + \tau\gamma\omega(L)\big)\|y^k - y^\star\|_{\gamma,\tau}^2 + \kappa\gamma^2\sigma_k^2,$$

and

$$r = \max\left(1 - \gamma\mu, 1 - \rho + \frac{\beta}{\kappa}, \frac{1}{1 + \tau\gamma\omega(L)}\right),$$

we have

$$\mathbb{E}_k V^{k+1} \leq rV^k.$$