STOCHASTIC PRIMAL-DUAL HYBRID GRADIENT ALGORITHM WITH ARBITRARY SAMPLING AND IMAGING APPLICATIONS*

ANTONIN CHAMBOLLE†, MATTHIAS J. EHRHARDT‡, PETER RICHTÁRIK§, AND CAROLA-BIBIANE SCHÖNLIEB‡

Abstract. We propose a stochastic extension of the primal-dual hybrid gradient algorithm studied by Chambolle and Pock in 2011 to solve saddle point problems that are separable in the dual variable. The analysis is carried out for general convex-concave saddle point problems and problems that are either partially smooth / strongly convex or fully smooth / strongly convex. We perform the analysis for arbitrary samplings of dual variables, and obtain known deterministic results as a special case. Several variants of our stochastic method significantly outperform the deterministic variant on a variety of imaging tasks.

Key words. convex optimization, primal-dual algorithms, saddle point problems, stochastic optimization, imaging

AMS subject classifications. 65D18, 65K10, 74S60, 90C25, 90C15, 92C55, 94A08

1. Introduction.

Many modern problems in a variety of disciplines (imaging, machine learning, statistics, etc.) can be formulated as convex optimization problems. Instead of solving the optimization problems directly, it is often advantageous to reformulate the problem as a saddle point problem. A very popular algorithm to solve such saddle point problems is the primal-dual hybrid gradient (PDHG)[1] algorithm [37, 21, 13, 36, 14, 15]. It has been used to solve a vast amount of state-of-the-art problems—to name a few examples in imaging: image denoising with the structure tensor [22], total generalized variation denoising [11], dynamic regularization [7], multi-modal medical imaging [27], multi-spectral medical imaging [43], computation of non-linear eigenfunctions [26], regularization with directional total generalized variation [29]. Its popularity stems from two facts: First, it is very simple and therefore easy to implement. Second, it involves only simple operations like matrix-vector multiplications and evaluations of proximal operators which are for many problems of interest simple and in closed-form or easy to compute iteratively, cf. e.g. [33]. However, for large problems that are encountered in many real world applications, even these simple operations might be still too costly to perform too often.

*Submitted to the editors 15/06/2017.

Funding: A. C. benefited from a support of the ANR, "EANOI" Project I1148 / ANR-12-IS01-0003 (joint with FWF). Part of this work was done while he was hosted in Churchill College and DAMTP, Centre for Mathematical Sciences, University of Cambridge, thanks to a support of the French Embassy in the UK and the Cantab Capital Institute for Mathematics of Information. M. J. E. and C.-B. S. acknowledge support from Leverhulme Trust project "Breaking the non-convexity barrier", EPSRC grant "EP/M00483X/1", EPSRC centre "EP/N014588/1", the Cantab Capital Institute for the Mathematics of Information, and from CHiPS (Horizon 2020 RISE project grant). Moreover, C.-B. S. is thankful for support by the Alan Turing Institute. P. R. acknowledges the support of EPSRC Fellowship in Mathematical Sciences "EP/N005538/1" entitled "Randomized algorithms for extreme convex optimization".

†CMAP, Ecole Polytechnique, CNRS, France (antonin.chambolle@cmap.polytechnique.fr).

‡Department for Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK (m.j.ehrhardt@damtp.cam.ac.uk, cbs31@cam.ac.uk).

§(i) Visual Computing Center & Extreme Computing Research Center, KAUST, Thuwal, Saudi Arabia; (ii) School of Mathematics, University of Edinburgh, Edinburgh, United Kingdom; (iii) The Alan Turing Institute, London, United Kingdom (peter.richtarik@kaust.edu.sa, peter.richtarik@ed.ac.uk).

[1]We follow the terminology of [14] and call the algorithm simply PDHG. It corresponds to PDHGMu and PDHGMp in [21].

We propose a stochastic extension of the PDHG for saddle point problems that are separable in the dual variable (cf. e.g. [18, 50, 52, 34]) where not all but only a few of these operations are performed in every iteration. Moreover, as in incremental optimization algorithms [47, 31, 10, 9, 8, 45, 19] over the course of the iterations we continuously build up information from previous iterations which reduces variance and thereby negative effects of stochasticity. Non-uniform samplings [40, 38, 50, 39, 2] have been proven very efficient for stochastic optimization. In this work we use the expected separable overapproximation framework of [38, 39, 41] to prove all statements for all non-trivial and iteration-independent samplings.

Related Work. The proposed algorithm can be seen as a generalization of the algorithm of [18, 52, 50] to arbitrary blocks and a much wider class of samplings. Moreover, in contrast to their results, our results generalize the deterministic case considered in [37, 13, 36, 15]. Fercoq and Bianchi [23] proposed a stochastic primal-dual algorithm with explicit gradient steps that allows for larger step-sizes by averaging over previous iterates, however, this comes at the cost of prohibitively large memory requirements. Similar memory issues are encountered by a primal-dual algorithm of [4]. It is related to forward-backward splitting [30] and averaged gradient descent [10, 20] and therefore suffers the same memory issues as the averaged gradient descent. Moreover, Valkonen proposed a stochastic primal-dual algorithm that can exploit partial strong convexity of the saddle point functional [48]. Randomized versions of the alternating direction method of multipliers are discussed for instance in [51, 25]. In contrast to other works on stochastic primal-dual algorithms [35, 49], our analysis is not based on Fejér monotonicity [16]. We therefore do not prove almost sure convergence of the sequence but prove a variety of convergence rates depending on strong convexity assumptions instead. For smooth / strongly-convex problems, our analysis implies that the variance of the algorithm converges to zero which we will show empirically to be a dividing factor between our work and [35].

As a word of warning, our contribution should not be mistaken by other "stochastic" primal-dual algorithms, where errors in the computation of matrix-vector products and evaluation of proximal operators are modeled by random variables, cf. e.g. [35, 16, 44]. In our work we deliberately choose to compute only a subset of a whole iteration to save computational cost. These two notations are related but are certainly not the same.

1.1. Contributions. We briefly mention the main contributions of our work.

Generalization of Deterministic Case. The proposed stochastic algorithm is a direct generalization of the deterministic setting [37, 13, 36, 14, 15]. In the degenerate case where in every iteration all computations are performed, our algorithm coincides with the original deterministic algorithm. Moreover, the same holds true for our analysis of the stochastic algorithm where we recover almost all deterministic statements (boundedness, convergence rates etc.) [13, 36] in this degenerate case. Therefore, the theorems for both the deterministic and the stochastic case can be combined by a single proof.

Better Rates. Our analysis extends the simple setting of [50] such that the strong convexity assumptions and the sampling do not have to be uniform. In the special case of uniform strong convexity and uniform sampling, the proven convergence rates are better than the ones proven in [50].

Arbitrary Sampling. The proposed algorithm is guaranteed to converge under a very general class of samplings [38, 39, 41] and thereby generalizes also the algorithm of [50] which has only been analyzed for two specific samplings. As long as the sampling

is independent and identically distributed over the iterations and all computations have non-zero probability to be carried out, the theory holds and the algorithm will converge with the proven convergence rates.

Acceleration. We propose an acceleration of the stochastic primal-dual algorithm which accelerates the convergence from $\mathcal{O}(1/K)$ to $\mathcal{O}(1/K^2)$ if parts of the saddle point functional are strongly convex and thereby results in a significantly faster algorithm.

Scaling Invariance. In the strongly convex case, we propose parameters for several serial samplings (uniform, importance, optimal), all based on the condition numbers of the problem and thereby independent of scaling.

2. General Problem. Let $\mathbb{X}, \mathbb{Y}_i, i = 1, \dots, n$ be real Hilbert spaces of any dimension and define the product space $\mathbb{Y} := \prod_{i=1}^n \mathbb{Y}_i$. For $y \in \mathbb{Y}$, we shall write $y = (y_1, y_2, \dots, y_n)$, where $y_i \in \mathbb{Y}_i$. Further, we consider the natural inner product on the product space $\mathbb{Y}$ given by $\langle y, z\rangle = \sum_{i=1}^n \langle y_i, z_i\rangle$, where $y_i, z_i \in \mathbb{Y}_i$. This inner product induces the norm $\|y\|^2 = \sum_{i=1}^n \|y_i\|^2$. Let $\mathbf{A} : \mathbb{X} \to \mathbb{Y}$ be a bounded linear operator. Due to the product space nature of $\mathbb{Y}$, we have $(\mathbf{A}x)_i = \mathbf{A}_i x$, where $\mathbf{A}_i : \mathbb{X} \to \mathbb{Y}_i$ are linear operators. The adjoint of $\mathbf{A}$ is given by $\mathbf{A}^* y = \sum_{i=1}^n \mathbf{A}_i^* y_i$. Moreover, let $f : \mathbb{Y} \to \mathbb{R}_\infty := \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{X} \to \mathbb{R}_\infty$ be convex functions. In particular, we assume that f is separable, i.e. $f(y) = \sum_{i=1}^n f_i(y_i)$.

Given the setup described above, we consider the optimization problem

$$\min_{x\in\mathbb{X}} \left\{ \Phi(x) := \sum_{i=1}^n f_i(\mathbf{A}_i x) + g(x) \right\}. \tag{1}$$

Instead of solving (1) directly, it is often desirable to reformulate the problem as a saddle point problem with the help of the Fenchel conjugate. If f is proper, convex, and lower semi-continuous, then $f(y) = f^{**}(y) = \sup_{z\in\mathbb{Y}} \langle z, y\rangle - f^*(z)$ where $f^* : \mathbb{Y} \to \mathbb{R} \cup \{-\infty, +\infty\}$, and $f^*(y) = \sum_{i=1}^n f_i^*(y_i)$ is the Fenchel conjugate of f (and f^{**} its biconjugate defined as the conjugate of the conjugate). Then solving (1) is equivalent to finding the primal part x of a solution to the saddle point problem (called a saddle point)

$$\min_{x\in\mathbb{X}} \sup_{y\in\mathbb{Y}} \left\{ \sum_{i=1}^n \langle \mathbf{A}_i x, y_i\rangle - f_i^*(y_i) + g(x) \right\}. \tag{2}$$

We will assume that the saddle point problem (2) has a solution. For conditions for existence and uniqueness, we refer the reader to [5]. A saddle point $(x^\sharp, y^\sharp) = (x^\sharp, y_1^\sharp, \dots, y_n^\sharp)$ satisfies the optimality conditions

$$\begin{aligned} \mathfrak{f}_i^\sharp &:= \mathbf{A}_i x^\sharp \in \partial f_i^*(y_i^\sharp), \quad \text{for } i = 1, \dots, n \\ \mathfrak{g}^\sharp &:= -\mathbf{A}^* y^\sharp \in \partial g(x^\sharp). \end{aligned}$$

An important notion in this work is *strong convexity*. A functional g is called μ_g-convex if $g - \frac{\mu_g}{2}\|\cdot\|^2$ is convex. In general, we assume that g is μ_g-convex, f_i^* is μ_i-convex with nonnegative strong convexity parameters $\mu_g, \mu_i \geq 0$. The convergence results in this contribution cover three different cases of regularity: i) no strong convexity $\mu_g, \mu_i = 0$, ii) semi-strong convexity $\mu_g > 0$ or $\mu_i > 0$ and iii) full strong convexity $\mu_g, \mu_i > 0$. For notational convenience we make use of the operators $\mathbf{M}_i := \mu_i \mathbf{I}, \mathbf{M}_g := \mu_g \mathbf{I}$ and $\mathbf{M}_f := \operatorname{diag}(\mathbf{M}_1, \dots, \mathbf{M}_n)$.

A very popular algorithm to solve the saddle point problem (2) is the Primal-Dual Hybrid Gradient (PDHG) algorithm [37, 21, 13, 36, 14, 15]. It reads (with extrapolation on y)

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_g^\tau \left(x^k - \tau \mathbf{A}^* \overline{y}^k\right) \\ y^{k+1} &= \operatorname{prox}_{f^*}^\sigma \left(y^k + \sigma \mathbf{A} x^{k+1}\right) \\ \overline{y}^{k+1} &= y^{k+1} + \theta(y^{k+1} - y^k)\,, \end{aligned}$$

where the *proximal operator (or proximity / resolvent operator)* is defined as

$$\operatorname{prox}_f^\tau (y) := \arg\min_{x \in \mathbb{X}} \left\{ \frac{1}{2}\|x - y\|_{\tau^{-1}}^2 + f(x) \right\}$$

and the weighted norm by $\|x\|_{\tau^{-1}}^2 = \langle \tau^{-1} x, x\rangle$. Its convergence is guaranteed if the step size parameters σ, τ are positive and satisfy $\sigma\tau\|\mathbf{A}\|^2 < 1$ and $\theta = 1$ [13]. Note that the definition of the proximal operator is well-defined for a *operator-valued* step size τ. In the case of a separable function f and with operator-valued step sizes the PDHG algorithm takes the form

$$x^{k+1} = \operatorname{prox}_g^{\mathbf{T}} \left(x^k - \mathbf{T}\mathbf{A}^* \overline{y}^k\right) \tag{3a}$$

$$y_i^{k+1} = \operatorname{prox}_{f_i^*}^{\mathbf{S}_i} \left(y_i^k + \mathbf{S}_i \mathbf{A}_i x^{k+1}\right)\,, \quad i = 1, \dots, n \tag{3b}$$

$$\overline{y}^{k+1} = y^{k+1} + \theta(y^{k+1} - y^k)\,. \tag{3c}$$

Here the step size parameters $\mathbf{S} = \operatorname{diag}(\mathbf{S}_1, \dots, \mathbf{S}_n)$ (a block diagonal operator) and $\mathbf{S}_1, \dots, \mathbf{S}_n$ and $\mathbf{T}$ are symmetric and positive definite. The algorithm is guaranteed to converge if $\|\mathbf{S}^{1/2}\mathbf{A}\mathbf{T}^{1/2}\| < 1$ and $\theta = 1$ [36].

3. Algorithm. In this work we extend the PDHG algorithm to a stochastic setting where in each iteration we update a random subset $\mathbb{S}$ of the dual variables (3b). This subset is sampled in an i.i.d. fashion from a fixed but otherwise *arbitrary* distribution, whence the name "arbitrary sampling". In order to guarantee convergence, it is necessary to assume that the sampling is "proper" [42, 39]. A sampling is proper if for each dual variable i we have $i \in \mathbb{S}$ with a positive probability $p_i > 0$. Examples of proper samplings include the *full sampling* where $\mathbb{S} = \{1, \dots, n\}$ with probability 1 and serial sampling where $\mathbb{S} = \{i\}$ is chosen with probability p_i. It is important to note that also other samplings are admissible. For instance for $n = 3$, consider the sampling that selects $\mathbb{S} = \{1, 2\}$ with probability $1/3$ and $\mathbb{S} = \{2, 3\}$ with probability $2/3$. Then the probabilities for the three blocks are $p_1 = 1/3$, $p_2 = 1$ and $p_3 = 2/3$ which makes it a proper sampling. However, if only $\mathbb{S} = \{1, 2\}$ is chosen with probability 1, then this sampling is not proper as the probability for the third block is zero: $p_3 = 0$.

The algorithm we propose is formalized as Algorithm 1. As in the original algorithm, the stepsize parameters $\mathbf{T}, \mathbf{S}_i$ have to be self-adjoint and positive definite operators for the updates to be well-defined. The extrapolation is performed with a scalar $\theta > 0$ and an operator $\mathbf{P} := \operatorname{diag}(p_1\mathbf{I}, \dots, p_n\mathbf{I})$ of probabilities p_i that an index is selected in each iteration.

REMARK 1. *Both, the primal and dual iterates x^{k+1} and y^{k+1} are random variables but only the dual iterate y^{k+1} depends on the sampling $\mathbb{S}^{k+1}$.*

REMARK 2 (Memory). *For memory efficiency, the algorithm can be coded in a different order, where the primal update follows the dual update. Our analysis obviously translates to this memory efficient algorithm.*

Algorithm 1 Stochastic Primal-Dual Hybrid Gradient algorithm (**SPDHG**). **Input:** primal and dual variable x^0, y^0, step length parameters $\mathbf{S} = \operatorname{diag}(\mathbf{S}_1, \dots, \mathbf{S}_n)$, $\mathbf{T}, \theta$, selection rule $k \mapsto \mathbb{S}^k$, number of iterations K. **Initialize:** $\overline{y}^0 = y^0$

for $k = 0, \dots, K-1$ **do**

$x^{k+1} = \operatorname{prox}_g^{\mathbf{T}}\left(x^k - \mathbf{T}\mathbf{A}^*\overline{y}^k\right)$

Select $\mathbb{S}^{k+1} \subset \{1, \dots, n\}$.

$$y_i^{k+1} = \begin{cases} \operatorname{prox}_{f_i^*}^{\mathbf{S}_i}\left(y_i^k + \mathbf{S}_i\mathbf{A}_i x^{k+1}\right) & \text{if } i \in \mathbb{S}^{k+1} \\ y_i^k & \text{else} \end{cases}$$

$\overline{y}^{k+1} = y^{k+1} + \theta\mathbf{P}^{-1}(y^{k+1} - y^k)$

end for

REMARK 3 (Computation). *Due to the sampling each iteration requires both* $\mathbf{A}_i$ *and* $\mathbf{A}_i^*$ *to be evaluated only for each selected index* $i \in \mathbb{S}^{k+1}$*. To see this, note that*

$$\mathbf{A}^*\overline{y}^{k+1} = \mathbf{A}^*y^{k+1} + \mathbf{A}^*\theta\mathbf{P}^{-1}(y^{k+1} - y^k) = \mathbf{A}^*y^k + \sum_{i \in \mathbb{S}^{k+1}} \left(1 + \frac{\theta}{p_i}\right) \mathbf{A}_i^*(y_i^{k+1} - y_i^k)\,.$$

$\mathbf{A}^*y^k$ *can be stored from the previous iteration (needs memory of the size of the primal variable* x*) and update requires the operators* $\mathbf{A}_i^*$ *to be evaluated only for* $i \in \mathbb{S}^{k+1}$*.*

4. General Convex Case. We first analyze the convergence of Algorithm 1 in the general convex case without making use of any strong convexity or smoothness assumptions. In order to analyze the convergence for the large class of samplings described in the previous section we make use of the *expected separable overapproximation (ESO)* inequality [39].

DEFINITION 4.1 (Expected Separable Overapproximation (ESO) [39]).

$$\mathbb{E}_{\mathbb{S}}\left\|\sum_{i \in \mathbb{S}} \mathbf{C}_i^* y_i\right\|^2 \leq \sum_{i=1}^n p_i v_i \|y_i\|^2\,. \tag{4}$$

Such parameters $\{v_i\}$ *are called* ESO parameters*.*

REMARK 4. *Note that for any operator* $\mathbf{C}^*$ *such parameters always exist but are obviously not unique. For the efficiency of the algorithm it is desirable to find ESO parameters such that* (4) *is as tight as possible; i.e., we want the parameters* $\{v_i\}$ *be small. As we shall see, the ESO parameters influence the choice of the aggregation parameter* θ*.*

The ESO inequality was first proposed by Richtárik and Takáč [42] to study parallel coordinate descent methods in the context of *uniform* samplings, which are samplings for which $p_i = p_j$ for all i, j. Improved bounds for ESO parameters were obtained in [24] and used in the context of accelerated coordinate descent. Qu et al. [39] perform an in-depth study of ESO parameters. The ESO inequality is also critical in the study mini-batch stochastic gradient descent with [28] or without [46] variance reduction.

We will frequently need to estimate the expected value of inner products which we will do by means of ESO parameters. Recall that we defined weighted norms as $\|x\|_{\mathbf{T}^{-1}}^2 := \langle \mathbf{T}^{-1}x, x\rangle$.

LEMMA 4.2. *Let y^k be defined as in Algorithm 1, v_i be some ESO parameters of $\mathbf{C}^* = [\mathbf{C}_1^*, \dots, \mathbf{C}_n^*]$ with $\mathbf{C}_i = \mathbf{S}_i^{1/2} p_i^{-1/2} \mathbf{A}_i \mathbf{T}^{1/2}$ and let $\gamma^2 \geq \max_i v_i$. Denote the expected value with respect to the sampling $\mathbb{S}^k$ by $\mathbb{E}^k$. Then for any $x \in \mathbb{X}$ and $c > 0$*

$$\mathbb{E}^k \langle x, \mathbf{A}^* \mathbf{P}^{-1} (y^k - y^{k-1}) \rangle \geq -\mathbb{E}^k \left\{ \frac{c}{2} \|x\|^2_{\mathbf{T}^{-1}} + \frac{\gamma^2}{2c} \|y^k - y^{k-1}\|^2_{(\mathbf{SP})^{-1}} \right\}.$$

EXAMPLE 1 (Full Sampling). *Let $\mathbb{S} = \{1, \dots, n\}$ with probability 1 such that $p_i = \mathbb{P}(i \in \mathbb{S}) = 1$. Then $v_i = \|\mathbf{S}^{1/2} \mathbf{A} \mathbf{T}^{1/2}\|^2$ are ESO parameters of $\mathbf{S}^{1/2} \mathbf{P}^{-1/2} \mathbf{A} \mathbf{T}^{1/2}$. Thus, the deterministic condition on convergence $\|\mathbf{S}^{1/2} \mathbf{A} \mathbf{T}^{1/2}\|^2 < 1$ implies that there exist some ESO parameters v_i (e.g. the above) that are uniformly bounded by γ^2 with $\gamma < 1$.*

EXAMPLE 2 (Serial Sampling). *Let $\mathbb{S} = \{i\}$ be chosen with probability $p_i > 0$. Then $v_i = \|\mathbf{S}_i^{1/2} p_i^{-1/2} \mathbf{A}_i \mathbf{T}^{1/2}\|^2$ are ESO parameters of $\mathbf{S}^{1/2} \mathbf{P}^{-1/2} \mathbf{A} \mathbf{T}^{1/2}$.*

For notational convenience we define[2] the function $h^* : \mathbb{Y} \to \mathbb{R}_\infty$ via $h^*(y) := \sum_{i=1}^n \left(\frac{1}{p_i} - 1\right) f_i^*(y_i)$ and denote the subgradient as $\mathfrak{h}^\sharp := \sum_{i=1}^n \left(\frac{1}{p_i} - 1\right) \mathfrak{f}_i^\sharp \in \partial h^*(y^\sharp)$. The Bregman distance of h^* is then given by

$$D^{\mathfrak{h}^\sharp}_{h^*}(y, y^\sharp) = \sum_{i=1}^n \left(\frac{1}{p_i} - 1\right) D^{\mathfrak{f}_i^\sharp}_{f_i^*}(y_i, y_i^\sharp). \tag{5}$$

THEOREM 4.3. *Let $(x^\sharp, y^\sharp) \in \mathbb{X} \times \mathbb{Y}$ be any saddle point, g, f_i be convex and the extrapolation parameter $\theta = 1$. Let the step sizes $\mathbf{T}, \mathbf{S}$ be chosen such that $0 < \gamma^2 < 1$ bounds some ESO parameters as defined by Lemma 4.2. Then, the iterates of Algorithm 1 satisfy the following assertions.*

1. *The sequence (x^K, y^K) is bounded in expectation in the sense that*

$$\mathbb{E} \left\{ \frac{1 - \gamma^2}{2} \|x^K - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2} \|y^K - y^\sharp\|^2_{(\mathbf{SP})^{-1}} + D^{\mathfrak{h}^\sharp}_{h^*}(y^K, y^\sharp) \right\} \leq c \tag{6}$$

where the constant is given by

$$c = \frac{1}{2} \|x^0 - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2} \|y^0 - y^\sharp\|^2_{(\mathbf{SP})^{-1}} + D^{\mathfrak{h}^\sharp}_{h^*}(y^0, y^\sharp). \tag{7}$$

2. *The Bregman distances between iterates (x^K, y^K) and a saddle point are almost surely summable, i.e. $\sum_{K=1}^\infty D^{\mathfrak{g}^\sharp}_g(x^K, x^\sharp), \sum_{K=1}^\infty D^{\mathfrak{f}^\sharp}_{f^*}(y^K, y^\sharp) < \infty$ almost surely. In particular, the Bregman distances converge to zero almost surely.*
3. *The ergodic sequence $(x_K, y_K) := \frac{1}{K} \sum_{k=1}^K (x^k, y^k)$ converges with rate $1/K$ in an expected Bregman sense to a saddle point, i.e.*

$$\mathbb{E} \left\{ D^{\mathfrak{g}^\sharp}_g(x_K, x^\sharp) + D^{\mathfrak{f}^\sharp}_{f^*}(y_K, y^\sharp) \right\} \leq \frac{c}{K} \tag{8}$$

where the constant is given by (7).

Proof of Theorem 4.3. The result of Lemma A.1 has to be adapted to the stochastic setting as the next iterations takes the value of $\hat{y}_i^{k+1}$ only with a certain probability.

[2] h^* is the convex conjugate of some function h, but this is not of importance in this manuscript.

First, for the Bregman distance of f^* we derive with the standard result of Lemma A.3

$$\begin{aligned}
D_{f^*}^{\mathfrak{f}^\sharp}(\hat{y}^{k+1}, y^\sharp) &= \sum_{i=1}^n \frac{1}{p_i}\mathbb{E}^{k+1} D_{f_i^*}^{\mathfrak{f}_i^\sharp}(y_i^{k+1}, y_i^\sharp) - \left(\frac{1}{p_i} - 1\right) D_{f_i^*}^{\mathfrak{f}_i^\sharp}(y_i^k, y_i^\sharp) \\
&= \sum_{i=1}^n \left(\frac{1}{p_i} - 1\right)\left(\mathbb{E}^{k+1} D_{f_i^*}^{\mathfrak{f}_i^\sharp}(y_i^{k+1}, y_i^\sharp) - D_{f_i^*}^{\mathfrak{f}_i^\sharp}(y_i^k, y_i^\sharp)\right) + \mathbb{E}^{k+1} D_{f^*}^{\mathfrak{f}^\sharp}(y^{k+1}, y^\sharp) \\
&= \mathbb{E}^{k+1} D_{h^*}^{\mathfrak{h}^\sharp}(y^{k+1}, y^\sharp) - D_{h^*}^{\mathfrak{h}^\sharp}(y^k, y^\sharp) + \mathbb{E}^{k+1} D_{f^*}^{\mathfrak{f}^\sharp}(y^{k+1}, y^\sharp)
\end{aligned}$$

where we used the Bregman distance of h^* given by (5). Using this result and again Lemma A.3 we can rewrite the estimate of Lemma A.1 as

$$\begin{aligned}
&\frac{1}{2}\|x^k - x^\sharp\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^k - y^\sharp\|_{(\mathbf{SP})^{-1}}^2 + D_{h^*}^{\mathfrak{h}^\sharp}(y^k, y^\sharp) \\
&\geq \mathbb{E}^{k+1}\left\{\frac{1}{2}\|x^{k+1} - x^\sharp\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^{k+1} - y^\sharp\|_{(\mathbf{SP})^{-1}}^2 + D_{h^*}^{\mathfrak{h}^\sharp}(y^{k+1}, y^\sharp) + D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp)\right. \\
&\quad + D_{f^*}^{\mathfrak{f}^\sharp}(y^{k+1}, y^\sharp) - \langle \mathbf{A}(x^{k+1} - x^\sharp), \mathbf{P}^{-1}y^{k+1} - (\mathbf{P}^{-1} - \mathbf{I})y^k - \overline{y}^k\rangle \\
&\quad \left. + \frac{1}{2}\|x^{k+1} - x^k\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^{k+1} - y^k\|_{(\mathbf{SP})^{-1}}^2\right\}.
\end{aligned}$$

where we have used the identity $\frac{1}{2}\|y^k - y^\sharp\|_{\mathbf{S}^{-1}}^2 + \frac{1}{2}\|y^k - y^\sharp\|_{\mathbf{S}^{-1}(\mathbf{P}^{-1}-\mathbf{I})}^2 = \frac{1}{2}\|y^k - y^\sharp\|_{(\mathbf{SP})^{-1}}^2$ to simplify the expression.

Plugging in the extrapolation $\overline{y}^k = y^k + \theta\mathbf{P}^{-1}(y^k - y^{k-1})$, $\theta = 1$, taking the expectations $\mathbb{E}^k, \mathbb{E}^{k-1}$ and denoting

$$\Delta^k := \mathbb{E}^k\mathbb{E}^{k-1}\left(\frac{1}{2}\|x^k - x^\sharp\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^k - y^\sharp\|_{(\mathbf{SP})^{-1}}^2 + D_{h^*}^{\mathfrak{h}^\sharp}(y^k, y^\sharp)\right)$$

leads to

$$\begin{aligned}
\Delta^k \geq \Delta^{k+1} + \mathbb{E}^{k+1}\mathbb{E}^k\mathbb{E}^{k-1}&\left\{\frac{1}{2}\|x^{k+1} - x^k\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^{k+1} - y^k\|_{(\mathbf{SP})^{-1}}^2\right. \\
&+ \langle x^{k+1} - x^k, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle + D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp) + D_{f^*}^{\mathfrak{f}^\sharp}(y^{k+1}, y^\sharp) \\
&\left. + \langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle - \langle x^{k+1} - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^{k+1} - y^k)\rangle\right\}.
\end{aligned} \tag{9}$$

Summing (9) over $k = 0, \ldots, K-1$ (note that $y^{-1} := y^0$) yields

$$\begin{aligned}
\Delta^0 \geq \Delta^K + \sum_{k=0}^{K-1} \mathbb{E}^{k+1}\mathbb{E}^k\mathbb{E}^{k-1}&\left\{\frac{1}{2}\|x^{k+1} - x^k\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^{k+1} - y^k\|_{(\mathbf{SP})^{-1}}^2\right. \\
&\left. + \langle x^{k+1} - x^k, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle + D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp) + D_{f^*}^{\mathfrak{f}^\sharp}(y^{k+1}, y^\sharp)\right\} \\
&- \mathbb{E}^K\mathbb{E}^{K-1}\langle x^K - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^K - y^{K-1})\rangle.
\end{aligned}$$

Moreover, estimating the inner products with Lemma 4.2

$$\mathbb{E}^k\langle x, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle \geq -\frac{\gamma^2}{2}\mathbb{E}^k\|x\|_{\mathbf{T}^{-1}}^2 - \frac{1}{2}\mathbb{E}^k\|y^k - y^{k-1}\|_{(\mathbf{SP})^{-1}}^2$$

and taking expectations with respect to $\mathbb{S}^1, \dots, \mathbb{S}^K$ (denoting this by $\mathbb{E}$) yields

$$\begin{aligned}
&\frac{1}{2}\|x^0 - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2}\|y^0 - y^\sharp\|^2_{(\mathbf{SP})^{-1}} + D^{\mathfrak{h}^\sharp}_{h^*}(y^0, y^\sharp) \\
&\geq \mathbb{E}\left\{\frac{1-\gamma^2}{2}\|x^K - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2}\|y^K - y^\sharp\|^2_{(\mathbf{SP})^{-1}} + D^{\mathfrak{h}^\sharp}_{h^*}(y^K, y^\sharp)\right\} \\
&\quad + \mathbb{E}\sum_{k=1}^{K}\left\{D^{\mathfrak{g}^\sharp}_{g}(x^k, x^\sharp) + D^{\mathfrak{f}^\sharp}_{f^*}(y^k, y^\sharp) + \frac{1-\gamma^2}{2}\|x^k - x^{k-1}\|^2_{\mathbf{T}^{-1}}\right\}.
\end{aligned} \tag{10}$$

All assertions of the theorem follow from inequality (10). It is easy to see that the sequence (x^K, y^K) is bounded in the sense of Item 1 as Bregman distances and norms are non-negative. For Item 2, note that it follows from (10) that $\mathbb{E}\sum_{k=1}^{\infty} D^{\mathfrak{g}^\sharp}_{g}(x^k, x^\sharp) < \infty$, $\mathbb{E}\sum_{k=1}^{\infty} D^{\mathfrak{f}^\sharp}_{f^*}(y^k, y^\sharp) < \infty$ which implies almost surely $\sum_{k=0}^{\infty} D^{\mathfrak{g}^\sharp}_{g}(x^k, x^\sharp) < \infty$, $\sum_{k=0}^{\infty} D^{\mathfrak{f}^\sharp}_{f^*}(y^k, y^\sharp) < \infty$ and thus that the Bregman distance between the iterates and any saddle point converges to zero almost surely. To see Item 3, note that Bregman distances of convex functions are convex in the first argument. Thus, dividing (10) by K yields

$$\mathbb{E}\Big\{D^{\mathfrak{g}^\sharp}_{g}(x_K, x^\sharp) + D^{\mathfrak{f}^\sharp}_{f^*}(y_K, y^\sharp)\Big\} \leq \frac{1}{K}\mathbb{E}\sum_{k=1}^{K}\Big\{D^{\mathfrak{g}^\sharp}_{g}(x^k, x^\sharp) + D^{\mathfrak{f}^\sharp}_{f^*}(y^k, y^\sharp)\Big\} \leq \frac{c}{K}$$

which was to be shown. ∎

5. Semi-Strongly Convex Case. In this section we propose two algorithms that converge as $\mathcal{O}(1/k^2)$ if either f_i^* or g is strongly convex.

Algorithm 2 Stochastic Primal-Dual Hybrid Gradient algorithm with acceleration on the primal variable (**PA-SPDHG**). The algorithm is only defined for scalar-valued primal step sizes, i.e. $\mathbf{T}^k = \tau^k \mathbf{I}$. **Input:** primal and dual variable x^0, y^0, step length parameters $\tau^0, \mathbf{S}^0$, selection rule $k \mapsto \mathbb{S}^k$, number of iterations K. **Initialize:** $\overline{y}^0 = y^0$

1: **for** $k = 0, \dots, K-1$ **do**
2: $\quad x^{k+1} = \operatorname{prox}^{\tau^k}_{g}\left(x^k - \tau^k \mathbf{A}^* \overline{y}^k\right)$
3: $\quad$ Select $\mathbb{S}^{k+1} \subset \{1, \dots, n\}$.
4: $\quad y_i^{k+1} = \begin{cases} \operatorname{prox}^{\mathbf{S}_i^k}_{f_i^*}\left(y_i^k + \mathbf{S}_i^k \mathbf{A}_i x^{k+1}\right) & \text{if } i \in \mathbb{S}^{k+1} \\ y_i^k & \text{else} \end{cases}$
5: $\quad \theta^k = (1 + 2\mu_g \tau^k)^{-1/2}, \quad \tau^{k+1} = \theta^k \tau^k, \quad \mathbf{S}^{k+1} = \mathbf{S}^k / \theta^k$
6: $\quad \overline{y}^{k+1} = y^{k+1} + \theta^k \mathbf{P}^{-1}(y^{k+1} - y^k)$
7: **end for**

THEOREM 5.1 (Dual Strong Convexity). *Let f_i^* be strongly convex with strong convexity constants $\mu_i > 0, i = 1, \dots, n$, the step sizes $\tilde{\sigma}^0, \mathbf{T}^0$ be chosen such that $0 < \gamma^2 \leq 1$ bounds some ESO parameters as defined by Lemma 4.2 and*

$$\tilde{\sigma}^0 < \min_i \frac{p_i}{2(1-p_i)}. \tag{11}$$

Let (x^K, y^K) be defined by Algorithm 3 and $\mathbf{Y}^k := (\mathbf{S}^k\mathbf{P})^{-1} + 2\mathbf{M}_f(\mathbf{P}^{-1} - \mathbf{I})$. Then there exists $\tilde{K} \in \mathbb{N}$ such that for all $K \geq \tilde{K}$ it holds

$$\mathbb{E}\|y^K - y^\sharp\|^2_{\mathbf{Y}^0} \leq \frac{2}{K^2}\Big\{\|x^0 - x^\sharp\|^2_{(\mathbf{T}^0)^{-1}} + \|y^0 - y^\sharp\|^2_{\mathbf{Y}^0}\Big\}.$$

Algorithm 3 Stochastic Primal-Dual Hybrid Gradient algorithm with acceleration on the dual variable (**DA-SPDHG**). The algorithm is only defined for scalar-valued dual step sizes, i.e. $\mathbf{S}_i^k = \sigma_i^k \mathbf{I}, i = 1, \dots, n$. **Input:** primal and dual variable x^0, y^0, step length parameters $\mathbf{T}^0, \tilde{\sigma}^0$, selection rule $k \mapsto \mathbb{S}^k$, number of iterations K. **Initialize:** $\overline{y}^0 = y^0$

1: **for** $k = 0, \dots, K-1$ **do**
2: $\quad x^{k+1} = \operatorname{prox}_g^{\mathbf{T}^k}\left(x^k - \mathbf{T}^k \mathbf{A}^* \overline{y}^k\right)$
3: $\quad$ Select $\mathbb{S}^{k+1} \subset \{1, \dots, n\}$.
4: $\quad \sigma_i^k = \frac{\tilde{\sigma}^k}{\mu_i[p_i - 2(1-p_i)\tilde{\sigma}^k]}, \quad i \in \mathbb{S}^{k+1}$
5: $\quad y_i^{k+1} = \begin{cases} \operatorname{prox}_{f_i^*}^{\sigma_i^k}\left(y_i^k + \sigma_i^k \mathbf{A}_i x^{k+1}\right) & \text{if } i \in \mathbb{S}^{k+1} \\ y_i^k & \text{else} \end{cases}$
6: $\quad \theta^k = (1 + 2\tilde{\sigma}^k)^{-1/2}, \quad \mathbf{T}^{k+1} = \mathbf{T}^k/\theta^k, \quad \tilde{\sigma}^{k+1} = \theta^k \tilde{\sigma}^k$
7: $\quad \overline{y}^{k+1} = y^{k+1} + \theta^k \mathbf{P}^{-1}(y^{k+1} - y^k)$
8: **end for**

REMARK 5. *As already noted in [13], $\tilde{K}$ is usually fairly small so that the estimate in Theorem 5.1 has practical relevance.*

REMARK 6. *In case of serial sampling and scalar primal step size $\tau^0 > 0$, the condition on the ESO parameters is equivalent to*

$$\tilde{\sigma}^0 \leq \min_i \frac{\gamma^2 \mu_i p_i^2}{\tau^0 \|\mathbf{A}_i\|^2 + 2\gamma^2 \mu_i p_i (1 - p_i)} .$$

In particular, it implies condition (11) on $\tilde{\sigma}^0$.

REMARK 7. *The convergence of Algorithm 2 with acceleration on the primal variable is similar to the deterministic case, cf. Appendix C.2 of [14], and omitted here for brevity. It converges with rate $\mathcal{O}(1/K^2)$ if there are ESO parameters that are bounded by $\gamma^2 < 1$.*

Proof of Theorem 5.1. The update on the step sizes in Algorithm 3 imply that

$$\|\cdot\|^2_{(\mathbf{T}^k)^{-1}} \geq (\theta^k)^{-1} \|\cdot\|^2_{(\mathbf{T}^{k+1})^{-1}}, \quad \text{and} \tag{12a}$$

$$\|\cdot\|^2_{(\mathbf{S}^k\mathbf{P})^{-1} + 2\mathbf{M}_f\mathbf{P}^{-1}} \geq (\theta^k)^{-1} \|\cdot\|^2_{(\mathbf{S}^{k+1}\mathbf{P})^{-1} + 2\mathbf{M}_f(\mathbf{P}^{-1} - I)} = (\theta^k)^{-1} \|\cdot\|^2_{\mathbf{Y}^{k+1}} . \tag{12b}$$

To see the latter, note the inequality is satisfied if

$$\theta^k \frac{1 + 2\mu_i \sigma_i^k}{p_i \mu_i \sigma_i^k} \geq \frac{1 + 2(1-p_i)\mu_i \sigma_i^{k+1}}{p_i \mu_i \sigma_i^{k+1}} \tag{13}$$

for all $i \in \{1, \dots, n\}$. With the auxiliary sequence

$$\tilde{\sigma}^k := \frac{p_i \mu_i \sigma_i^k}{1 + 2(1-p_i)\mu_i \sigma_i^k}, \quad \sigma_i^k = \frac{\tilde{\sigma}^k}{\mu_i[p_i - 2(1-p_i)\tilde{\sigma}^k]}$$

inequality (13) is satisfied as soon as

$$\theta^k \frac{1 + 2\tilde{\sigma}^k}{\tilde{\sigma}^k} \geq \frac{1}{\tilde{\sigma}^{k+1}} . \tag{14}$$

Note that the transformation from $\tilde{\sigma}^k$ to σ_i^k is well-defined if $\tilde{\sigma}^k < \min_i \frac{p_i}{2(1-p_i)}$. Under the additional assumption $\tilde{\sigma}^{k+1} = \theta^k \tilde{\sigma}^k$, (14) is solved with equality by $\theta^k = (1 + 2\tilde{\sigma}^k)^{-1/2}$. Moreover, the sequence of dual step sizes satisfies

$$\sigma_i^{k+1} = \frac{\theta^k \sigma_i^k}{1 + 2(1-\theta^k)(1-p_i)\mu_i \sigma_i^k} \leq \theta^k \sigma_i^k$$

which shows that the step size condition of Lemma 4.2 holds if it holds for the initial step size parameters.

For the actual proof of the theorem, note that the inequalities (12) imply

$$\begin{aligned} &\mathbb{E}\Big\{ \frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}} + \frac{1}{2}\|y^{k+1} - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1} + 2\mathbf{M}_f \mathbf{P}^{-1}} \\ &- \langle x^{k+1} - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^{k+1} - y^k)\rangle \Big\} \geq (\theta^k)^{-1}\mathbb{E}\Delta^{k+1} \end{aligned} \tag{15}$$

with $\Delta^k := \frac{1}{2}\|x^k - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{\mathbf{Y}^k} - \theta^{k-1}\langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1}\rangle$. Moreover, combining Lemma A.2 and (15) yields

$$\mathbb{E}\Big\{ \Delta^k + \frac{1}{2}\|y^k - y^{k-1}\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}} \Big\} \geq (\theta^k)^{-1}\mathbb{E}\Big\{ \Delta^{k+1} + \frac{1}{2}\|y^{k+1} - y^k\|^2_{(\mathbf{S}^{k+1}\mathbf{P})^{-1}} \Big\}.$$

where we used that $\mathbf{S}^{k+1} \leq \theta^k \mathbf{S}^k$, $\gamma^2(\theta^{k-1})^2 \leq 1$. Using this inequality recursively, $y^{-1} := y^0$, we arrive at

$$\begin{aligned} &\prod_{k=0}^{K-1} \theta^k \Big\{ \frac{1}{2}\|x^0 - x^\sharp\|^2_{(\mathbf{T}^0)^{-1}} + \frac{1}{2}\|y^0 - y^\sharp\|^2_{\mathbf{Y}^0} \Big\} \\ &\geq \mathbb{E}\Big\{ \frac{1}{2}\|x^K - x^\sharp\|^2_{(\mathbf{T}^K)^{-1}} + \frac{1}{2}\|y^K - y^\sharp\|^2_{\mathbf{Y}^K} - \langle x^K - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^K - y^{K-1})\rangle \\ &\quad + \frac{1}{2}\|y^K - y^{K-1}\|^2_{(\mathbf{S}^K\mathbf{P})^{-1}} \Big\} \\ &\geq \mathbb{E}\Big\{ \frac{1-\gamma}{2}\|x^K - x^\sharp\|^2_{(\mathbf{T}^K)^{-1}} + \frac{1}{2}\|y^K - y^\sharp\|^2_{\mathbf{Y}^K} + \frac{1-\gamma}{2}\|y^K - y^{K-1}\|^2_{(\mathbf{S}^K\mathbf{P})^{-1}} \Big\} \\ &\geq \frac{\tilde{\sigma}^0}{\tilde{\sigma}^K}\frac{1}{2}\mathbb{E}\|y^K - y^\sharp\|^2_{\mathbf{Y}^0} \end{aligned} \tag{16}$$

where the inner product for the second inequality is estimated by Lemma 4.2. The third inequality is due to the non-negativity of norms and $\|\cdot\|^2_{\mathbf{Y}^k} = \frac{1}{\tilde{\sigma}^k}\|\cdot\|^2_{\mathbf{M}_f} = \frac{\tilde{\sigma}^0}{\tilde{\sigma}^k}\|\cdot\|^2_{\mathbf{Y}^0}$ which holds by the definition of $\tilde{\sigma}^k$. Moreover, by the definition of $\theta^k = \frac{\tilde{\sigma}^{k+1}}{\tilde{\sigma}^k}$ (16) can be further simplified to

$$\mathbb{E}\|y^K - y^\sharp\|^2_{\mathbf{Y}^0} \leq \left(\frac{\tilde{\sigma}^K}{\tilde{\sigma}^0}\right)^2 \Big\{ \|x^0 - x^\sharp\|^2_{(\mathbf{T}^0)^{-1}} + \|y^0 - y^\sharp\|^2_{\mathbf{Y}^0} \Big\}.$$

Finally, the assertion follows by Corollary 1 of [13]. ∎

6. Strongly Convex Case. If both f_i^* and g are strongly convex, we may find step size parameters such that the algorithm Algorithm 1 converges linearly.

THEOREM 6.1. *Let $(x^\sharp, y^\sharp) \in \mathbb{X} \times \mathbb{Y}$ be a saddle point and g, f_i^* be strongly convex with constants $\mu_g, \mu_i > 0, i = 1, \ldots, n$. Let the step sizes $\mathbf{S}, \mathbf{T}, 0 < \theta < 1$ be chosen such that γ^2 bounds some ESO parameters as defined by Lemma 4.2, $\gamma^2\theta < 1$ and*

$$\theta(\mathbf{I} + 2\mathbf{M}_g\mathbf{T}) \geq \mathbf{I}, \qquad \text{and} \qquad \theta(\mathbf{I} + 2\mu_i\mathbf{S}_i) \geq \mathbf{I} + 2(1-p_i)\mu_i\mathbf{S}_i \quad i = 1, \ldots, n \tag{17}$$

in a positive semidefinite sense for operators. Let $\mathbf{X} := \mathbf{T}^{-1} + 2\mathbf{M}_g$, $\mathbf{Y} := (\mathbf{S}^{-1} + 2\mathbf{M}_f)\mathbf{P}^{-1}$. *Then the iterates of Algorithm 1 converge linearly to the saddle point, i.e.*

$$\mathbb{E}\Big\{(1-\gamma^2\theta)\|x^K - x^\sharp\|_{\mathbf{X}}^2 + \|y^K - y^\sharp\|_{\mathbf{Y}}^2\Big\} \leq \theta^K\Big\{\|x^0 - x^\sharp\|_{\mathbf{X}}^2 + \|y^0 - y^\sharp\|_{\mathbf{Y}}^2\Big\}.$$

Proof. The requirements (17) on the step sizes $\mathbf{S}, \mathbf{T}$ and θ imply $\|\cdot\|_{\mathbf{X}}^2 \geq \theta^{-1}\|\cdot\|_{\mathbf{T}^{-1}}^2$ and $\|\cdot\|_{\mathbf{Y}}^2 \geq \theta^{-1}\|\cdot\|_{(\mathbf{SP})^{-1}+2\mathbf{M}_f(\mathbf{P}^{-1}-\mathbf{I})}^2$. Thus, we directly get

$$\begin{aligned}\theta\,\mathbb{E}\Delta^k \geq \mathbb{E}\Bigg\{&\frac{1}{2}\|x^k - x^\sharp\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^k - y^\sharp\|_{(\mathbf{SP})^{-1}+2\mathbf{M}_f(\mathbf{P}^{-1}-\mathbf{I})}^2 \\ &- \theta\langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle\Bigg\}.\end{aligned} \tag{18}$$

where we denoted $\Delta^k := \frac{1}{2}\|x^k - x^\sharp\|_{\mathbf{X}}^2 + \frac{1}{2}\|y^k - y^\sharp\|_{\mathbf{Y}}^2 - \langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle$. Combining (18) and Lemma A.2 with constant step sizes yields

$$\theta\,\mathbb{E}\Delta^k \geq \mathbb{E}\left\{\Delta^{k+1} + \frac{1}{2}\|y^{k+1} - y^k\|_{(\mathbf{SP})^{-1}}^2 - \frac{\theta^2\gamma^2}{2}\|y^k - y^{k-1}\|_{(\mathbf{SP})^{-1}}^2\right\}.$$

Multiplying both sides by $\theta^{-(k+1)}$ and summing over $k = 0, \ldots, K-1$ yields

$$\begin{aligned}\Delta^0 &\geq \theta^{-K}\mathbb{E}\Big\{\Delta^K + \frac{1}{2}\|y^K - y^{K-1}\|_{(\mathbf{SP})^{-1}}^2\Big\} + \frac{1-\gamma^2\theta}{2}\mathbb{E}\sum_{k=1}^{K-1}\theta^{-k}\|y^k - y^{k-1}\|_{(\mathbf{SP})^{-1}}^2 \\ &\geq \theta^{-K}\mathbb{E}\Bigg\{\frac{1}{2}\|x^K - x^\sharp\|_{\mathbf{X}}^2 + \frac{1}{2}\|y^K - y^\sharp\|_{\mathbf{Y}}^2 + \frac{1}{2}\|y^K - y^{K-1}\|_{(\mathbf{SP})^{-1}}^2 \\ &\qquad - \langle x^K - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^K - y^{K-1})\rangle\Bigg\} \\ &\geq \theta^{-K}\mathbb{E}\left\{\frac{1}{2}\|x^K - x^\sharp\|_{\mathbf{X}}^2 - \frac{\gamma^2}{2}\|x^K - x^\sharp\|_{\mathbf{T}^{-1}}^2 + \frac{1}{2}\|y^K - y^\sharp\|_{\mathbf{Y}}^2\right\} \\ &\geq \theta^{-K}\mathbb{E}\left\{\frac{1-\gamma^2\theta}{2}\|x^K - x^\sharp\|_{\mathbf{X}}^2 + \frac{1}{2}\|y^K - y^\sharp\|_{\mathbf{Y}}^2\right\}\end{aligned}$$

where we used again Lemma 4.2 and the non-negativity of norms for the second inequality. Thus, the assertion is proven. ∎

6.1. Invariant Analysis. A desirable property of an algorithm is scaling invariance, i.e. the algorithm behaves independent of the scale of all variables and unknowns. If we rewrite problem (2) in terms of the scaled primal variable $\overline{x} := \alpha x$ and dual variables $\overline{y}_i := \beta_i y_i$, then the corresponding operators $\overline{\mathbf{A}}_i := \mathbf{A}_i/(\alpha\beta_i)$ have norm $\|\overline{\mathbf{A}}_i\| = \|\mathbf{A}_i\|/(\alpha\beta_i)$, the function $\overline{g}(\overline{x}) := g(\overline{x}/\alpha)$ is $\overline{\mu}_g := \mu_g/\alpha^2$ strongly convex and the functions $\overline{f}_i^*(\overline{y}_i) := f_i^*(\overline{y}_i/\beta_i)$ are $\overline{\mu}_i := \mu_i/\beta_i^2$ strongly convex. Thus the condition numbers $\kappa_i := \|\mathbf{A}_i\|^2/(\mu_g\mu_i)$

$$\overline{\kappa}_i = \frac{\|\overline{\mathbf{A}}_i\|^2}{\overline{\mu}_g\overline{\mu}_i} = \frac{\frac{1}{(\alpha\beta_i)^2}\|\mathbf{A}_i\|^2}{\frac{1}{\alpha^2}\mu_g\frac{1}{\beta_i^2}\mu_i} = \frac{\|\mathbf{A}_i\|^2}{\mu_g\mu_i} = \kappa_i$$

are scaling invariant.

6.2. Scalar Parameters for Serial Sampling. This analysis is to optimize the convergence rate θ of Theorem 6.1 for three different serial sampling options where exactly one block is chosen in each iteration. Other sampling strategies, including multi-block, parallel, etc. [39] will be subject of future work. For the ease of exposition we assume that $\mathbf{T} = \tau\mathbf{I}, \mathbf{S}_i = \sigma_i\mathbf{I}, i = 1, \dots, n$ are all scalar in the following. With $\bar{\sigma}_i := \sigma_i\mu_i$ and $\bar{\tau} := \tau\mu_g$, the conditions on the step size (17) become

$$\theta \geq \frac{1}{1+2\bar{\tau}}\,, \quad \theta \geq \max_i 1 - 2\frac{\bar{\sigma}_i p_i}{1+2\bar{\sigma}_i}\,, \quad \text{and} \quad \max_i \bar{\tau}\bar{\sigma}_i\kappa_i\theta \leq \rho^2 p_i \tag{19}$$

for some $\rho < 1$. The last condition arises from the ESO parameters of serial sampling. Finding optimal parameters is equivalent to equating the above inequalities. Note that the first two conditions (with equality) are equivalent to $\theta\bar{\tau} = (1-\theta)/2$ and $\bar{\sigma}_i = \frac{1-\theta}{2(p_i-(1-\theta))}$. With these choices, the third condition in (19) reads

$$(1-\theta)^2\tilde{\kappa} \leq 4p_i(p_i - (1-\theta)) \qquad i = 1, \dots, n\,. \tag{20}$$

where we denote $\tilde{\kappa} = \kappa/\rho^2$. It follows from (20) that for any $i = 1, \dots, n$ it holds

$$\theta \geq 1 - \frac{2p_i}{1+\sqrt{1+\tilde{\kappa}_i}} \tag{21}$$

1. **Serial uniform sampling:** We first consider uniform sampling, i.e. every block is sampled with the same probability $p_i = 1/n$. Then it is easy to see that the smallest achievable rate is given by

   $$\theta_{\text{uniform}} = 1 - \frac{2}{n + n\sqrt{1+\tilde{\kappa}_{\max}}} \tag{22}$$

   where $\tilde{\kappa}_{\max} := \max_j \tilde{\kappa}_j$ and the step sizes are then

   $$\sigma_i = \frac{\mu_i^{-1}}{\sqrt{1+\tilde{\kappa}_{\max}} - 1}\,, \quad \tau = \frac{\mu_g^{-1}}{n - 2 + n\sqrt{1+\tilde{\kappa}_{\max}}}\,.$$

2. **Serial importance sampling:** Instead of uniform sampling we may sample "important blocks" more often, i.e. we sample every block with a probability proportional to the square root of its condition number $p_i = \kappa_i^{1/2}/\sum_j \kappa_j^{1/2}$. Then the smallest rate that achieves (21) is given by

   $$\theta_{\text{importance}} = 1 - \frac{2\nu}{\sum_{j=1}^n \sqrt{\tilde{\kappa}_j}} \tag{23}$$

   and with $\nu := \frac{\sqrt{\tilde{\kappa}_{\min}}}{1+\sqrt{1+\tilde{\kappa}_{\min}}}$, $\tilde{\kappa}_{\min} := \min_j \tilde{\kappa}_j$ and the step sizes are

   $$\sigma_i = \frac{\nu\mu_i^{-1}}{\sqrt{\tilde{\kappa}_i} - 2\nu}\,, \quad \tau = \frac{\nu\mu_g^{-1}}{\sum_{j=1}^n \sqrt{\tilde{\kappa}_j} - 2\nu}\,.$$

3. **Serial optimal sampling:** Instead of a predefined probability we will seek for an "optimal sampling" that minimizes the linear convergence rate θ. The optimal sampling can be found by equating condition (21) for $i = 1, \dots, n$

   $$\theta\left(1 + \sqrt{1+\tilde{\kappa}_i}\right) = 1 + \sqrt{1+\tilde{\kappa}_i} - 2p_i\,. \tag{24}$$

Summing (24) from 1 to n and using that for serial sampling $\sum_{i=1}^n p_i = 1$ leads to

$$\theta_{\text{optimal}} = 1 - \frac{2}{n + \sum_{j=1}^n \sqrt{1+\tilde{\kappa}_j}} \tag{25}$$

with step size parameters

$$\sigma_i = \frac{\mu_i^{-1}}{\sqrt{1+\tilde{\kappa}_i} - 1}\,, \quad \tau = \frac{\mu_g^{-1}}{n - 2 + \sum_{j=1}^n \sqrt{1+\tilde{\kappa}_j}}$$

and probabilities

$$p_i = \frac{1+\sqrt{1+\tilde{\kappa}_i}}{n + \sum_{j=1}^n \sqrt{1+\tilde{\kappa}_j}}\,.$$

REMARK 8 (Minibatches). *All arguments above can easily be extended to samplings where at each iteration not only one but a predefined number m of blocks are chosen.*

REMARK 9 (Better Sampling). *It is easy to see that optimal sampling is better than uniform sampling: if all condition numbers are the same, then the rates for uniform sampling (22) and optimal sampling (25) are equal but if they are not, then the rate of optimal sampling is strictly smaller and thus better.*

Moreover, optimal sampling is better than importance sampling. To see this, note that due to the monotonicity of $\frac{\sqrt{x}}{1+\sqrt{1+x}}$ we get

$$\begin{aligned}\theta_{\text{importance}} &= 1 - \frac{2}{\left(1+\sqrt{1+\tilde{\kappa}_{\min}}\right)\sum_{j=1}^n \frac{\sqrt{\tilde{\kappa}_j}}{\sqrt{\tilde{\kappa}_{\min}}}} \\ &\geq 1 - \frac{2}{\left(1+\sqrt{1+\tilde{\kappa}_{\min}}\right)\sum_{j=1}^n \frac{1+\sqrt{1+\tilde{\kappa}_j}}{1+\sqrt{1+\tilde{\kappa}_{\min}}}} = \theta_{\text{optimal}}\,.\end{aligned}$$

REMARK 10 (Comparison to Zhang and Xiao [50]). *The algorithm of Zhang and Xiao [50] is (almost[3]) a special case of the proposed algorithm where each block is picked with probability $p_i = 1/n$. Here m denotes the size of each block to be processed at every iteration and n the number of blocks. Moreover, they only consider the strongly convex case where g is μ_g-strongly convex and all f_i^* are μ_f-strongly convex. Then with R being the largest norm of the rows in $\mathbf{A}$ they achieve*

$$\theta = 1 - \frac{1}{n + n\frac{\sqrt{m}R}{\sqrt{\mu_g \mu_f}}}\,.$$

If the minibatch size is $m = 1$, the blocks are chosen to be single rows and the probabilities are uniform, then their rate is worse than ours

$$\theta = 1 - \frac{1}{n + n\sqrt{\kappa_{\max}}} \geq 1 - \frac{2}{2n + n\sqrt{\tilde{\kappa}_{\max}}} \geq 1 - \frac{2}{2n + n(\sqrt{1+\tilde{\kappa}_{\max}} - 1)} = \theta_{\text{uniform}}$$

for any $\rho \geq \frac{1}{2}$. For $m > 1$, the rates differ even more as the condition numbers are conservatively estimated. Similarly, the rates can be improved by non-uniform sampling if the row norms are not equal.

[3] In contrast to our work, they have an extrapolation on both primal and dual variables. However, both extrapolations are related as our extrapolation factor is the product of their extrapolation factors.

7. Numerical Results. All numerical examples are implemented in python using numpy and the operator discretization library (ODL) [1]. The python code and all example data will be made available upon acceptance of this manuscript.

7.1. Non-Strongly Convex PET Reconstruction. In this example we consider positron emission tomography (PET) reconstruction with a total variation (TV) prior. The goal in PET imaging is to reconstruct the distribution of a radioactive tracer from its line integrals [32]. Let $\mathbb{X} = \mathbb{R}^{d_1 \times d_2}(d_1, d_2 = 250)$ be the space of tracer distributions (images) and $\mathbb{Y}_i = \mathbb{R}^{|B_i|}$ the data spaces where $B_i \subset \{1, \dots, N\}, N = 250 \times 354$ (250 views around the object) are subsets of indices with $B_i \cap B_j = \emptyset$ if $i \neq j$ and $\cup_{i=1}^n B_i = \{1, \dots, n\}$. All samplings in this example divide the views equidistantly. It is standard that PET reconstruction can be posed as the optimization problem

$$\min_{x \in \mathbb{X}} \left\{ \sum_{i=1}^{n} f_i(\mathbf{A}_i x) + g(x) \right\} \tag{26}$$

where the data fidelity term is given by the Kullback–Leibler divergence

$$f_i(y) = \begin{cases} \sum_{j \in B_i} y_j + r_j - b_j + b_j \log\left(\frac{b_j}{y_j + r_j}\right) & \text{if } y_j + r_j > 0, j \in B_i \\ \infty & \text{else,} \end{cases} \tag{27}$$

$0 \log 0 := 0$. The operator $\mathbf{A}$ is a scaled X-ray transform where in each of 250 directions 354 line integrals are computed. The prior is the TV of x with non-negativity constraint, i.e. $g(x) = \alpha \|\nabla x\|_{1,2} + \chi_{\geq 0}(x)$, with regularization parameter $\alpha = 4$ and the gradient operator $\nabla x \in \mathbb{R}^{d_1 \times d_2 \times 2}$ is discretized by forward differences, cf. [12] for details. The $1,2$-norm of these gradients is defined as $\|x\|_{1,2} := \sum_{i=1}^{d_1} \sum_{j=1}^{d_2} \sqrt{(\nabla_1 x_{i,j})^2 + (\nabla_2 x_{i,j})^2}$. The Fenchel conjugate of the Kullback–Leibler divergence (27) is

$$f_i^*(z) = \sum_{j \in B_i} \begin{cases} -z_j r_j - b_j \log(1 - z_j) & \text{if } z_j \leq 1 \text{ and } (b_j = 0 \text{ or } z_j < 1), j \in B_i \\ \infty & \text{else} \end{cases}, \tag{28}$$

its proximity operator given by

$$[\text{prox}_{f_i^*}^{\sigma}(z)]_j = \frac{1}{2}\left[z_j + 1 + \sigma r_j - \sqrt{(z_j - 1 + \sigma r_j)^2 + 4\sigma b_j} \right].$$

The proximal operator for g is approximated with 5 iterations of the fast gradient projection method (FGP) [6] with a warm start applied to the dual problem.

Parameters. In this experiment we choose $\gamma = 0.99$, $\theta = 1$ and all samplings are uniform, i.e. $p = 1/n$. The number of subsets varies between $n = 1$ (deterministic case), 50 and 250. The other step size parameters are chosen as

- PDHG and Pesquet and Repetti: $\sigma_i = \gamma / \|\mathbf{A}\|$, $\tau = \gamma / \|\mathbf{A}\|$
- SPDHG: $\sigma_i = \gamma / \|\mathbf{A}_i\|$, $\tau = \gamma / (n \max_i \|\mathbf{A}_i\|)$

Results. Some example reconstructed images are found in Figure 1 and quantitative results in Figures 2 and 3. It can be seen from the reconstructed images after 3 epochs in Figure 1 that both stochastic algorithms are much faster than the deterministic PDHG. The statements of Theorem 4.3 are numerically validated in Figure 2 which shows that the distance to a saddle point is bounded and that the ergodic

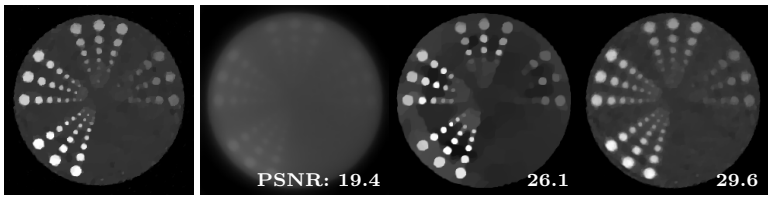

FIGURE 1. *PET reconstruction with TV prior solved as a non-strongly convex problem. Results after 3 epochs with uniform sampling of 250 subsets.* ***From left to right:*** *approximate primal part of saddle point, PDHG, Pesquet and Repetti [35] and SPDHG. With the same number of operator evaluations both stochastic algorithms make much more progress towards the saddle point.*

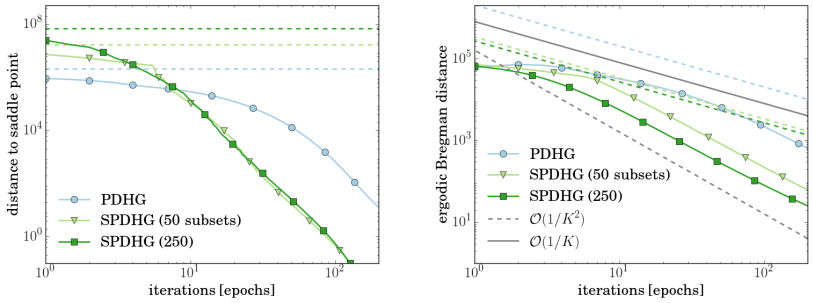

FIGURE 2. *Numerical evaluations (solid line) of the results in Theorem 4.3. Both the distance to a saddle point (6)* ***(left)*** *and the ergodic Bregman distance (8)* ***(right)*** *behave as predicted by the analysis (dashed line), however, the bounds are not sharp for this example.*

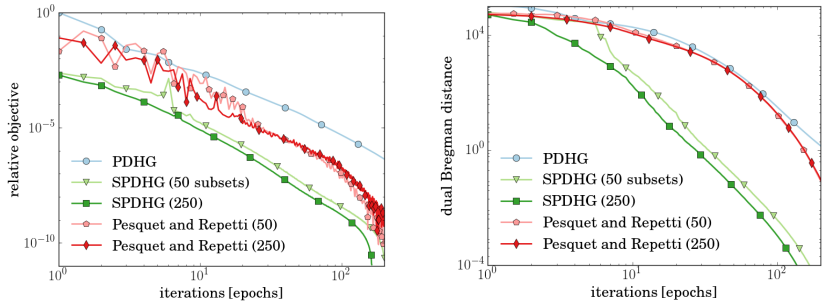

FIGURE 3. *Comparison of PDHG, SPDHG and the algorithm of Pesquet and Repetti [35] in terms of relative objective* $\frac{\Phi(x^K)-\Phi^\sharp}{\Phi(x^0)-\Phi^\sharp}$ ***(left)*** *and dual Bregman distance* $D^{\mathfrak{f}^\sharp}_{f^*}(y^K, y^\sharp)$ ***(right)****. Both graphs agree that the proposed algorithm converges faster than the algorithm of Pesquet and Repetti. Moreover, the graph of the relative objective function values* ***(left)*** *indicates that the proposed algorithm has also a much smaller variance compared to the algorithm of Pesquet and Repetti.*

Bregman distance converges with rate $1/k$. All five algorithms are compared quantitatively in Figure 3. In both, objective function value and dual Bregman distances, the stochastic algorithms are faster with SPDHG outperforming the algorithm of Pesquet and Repetti. In addition, two other observations can be made. First, SPDHG has empirically less variance than the algorithm of Pesquet and Repetti seen by the smoothness of the curves. Second, the stochastic algorithms differ a lot in terms of the convergence speed of the dual Bregman distance.

7.2. TV denoising with Gaussian Noise (Primal Acceleration). In the second example we consider denoising of an image that is degraded by Gaussian noise

with the help of the anisotropic TV. This can be achieved by solving the optimization problem

$$\min_{x\in\mathbb{X}} \left\{ \frac{1}{2\alpha}\|x-b\|_2^2 + \sum_{i=1}^{2} \|\nabla_i x\|_1 \right\}$$

where $\mathbb{X} = \mathbb{R}^{d_1\times d_2}$, the data fit is the squared Euclidean norm and the prior the (anisotropic) TV. The gradient is again discretized by forward differences, cf. e.g. [12]. Instead of the isotropic TV as in the previous example we consider here the anisotropic version as it is separable in the direction of the gradient. Dualizing the anisotropic TV yields the saddle point problem

$$\min_{x\in\mathbb{X}} \sup_{y\in\mathbb{Y}} \left\{ \sum_{i=1}^{2} \langle y_i, \nabla_i x\rangle - \imath_B(y_i) + \frac{1}{2\alpha}\|x-b\|_2^2 \right\}.$$

where $\mathbb{Y} = \mathbb{X}^2$ and $\imath_B$ is the characteristic function of the unit ball with respect to the ∞-norm, i.e.

$$\imath_B(y) := \begin{cases} 0 & \text{if } y_{i,j} \in [-1,1] \text{ for all } i,j \\ \infty & \text{else} \end{cases}. \tag{29}$$

For the primal-dual algorithms we need the proximal operators of the characteristic function and the squared 2-norm. These are given by the point-wise projection onto the unit interval $\operatorname{prox}^{\sigma}_{\imath_B}(z)_{i,j} = \frac{z_{i,j}}{\max(1,|z_{i,j}|)}$ and a shifted scaling $\operatorname{prox}^{\sigma}_{\frac{1}{2\alpha}\|\cdot-b\|_2^2}(z) = \frac{1}{\alpha+\sigma}(\alpha z + \sigma b)$. The regularization parameter is chosen to be $\alpha = 0.12$.

Parameters. In this experiment we choose $\gamma = 1$ and the sampling to be uniform, i.e. $p = 1/n$. The number of subsets are either $n = 1$ in the deterministic case or $n = 2$ in the stochastic case. The (initial) step size parameters are chosen as

- PDHG: $\sigma_i = 1/\|\mathbf{A}\|$, $\tau = 1/\|\mathbf{A}\|$
- PA-PDHG: $\sigma_i^0 = 1/\|\mathbf{A}\|$, $\tau^0 = 1/\|\mathbf{A}\|$
- SPDHG: $\sigma_i = 1/\|\mathbf{A}_i\|$, $\tau = 1/(n\max_i \|\mathbf{A}_i\|)$
- PA-SPDHG: $\sigma_i^0 = 1/\|\mathbf{A}_i\|$, $\tau^0 = 1/(n\max_i \|\mathbf{A}_i\|)$

For the accelerated version, the step sizes vary with the iteration with the primal step size getting smaller and the dual step size getting larger. The extrapolation factor is chosen to be $\theta = 1$ for the non-accelerated algorithms and converges to one for the accelerated versions.

Results. By visual assessment of the denoised images in Figure 4, it is easy to conclude that the accelerated algorithms are much faster than the non-accelerated version. Moreover, it can be seen that the stochastic variant of the accelerated algorithm is even faster than the deterministic version as the the sky is more uniform.

The quantitative results in Figure 5 confirm the visual conclusions. In addition, they show that the accelerated SPDHG indeed converges as $1/K^2$ in the norm of the primal part.

7.3. Huber-TV Deblurring with Unknown Boundary (Dual Acceleration). In the third example we consider the deblurring of an image with known convolution kernel but we do not assume to have knowledge about the boundary of the image [3]. The latter condition is very natural as no artificial boundary condition (zero, constant, periodic etc) will be met in a practical setting. Following the mathematical model of [3] the forward operator is modeled as $\mathbf{A} : \mathbb{X} \to \mathbb{R}^N, , \mathbb{X} =$

FIGURE 4. *Results for TV denoising after 20 epochs with uniform sampling.* ***Top:*** *noisy input data and approximate primal part of saddle point after 5000 PDHG iterations.* ***Bottom:*** *From left to right: PDHG, primal accelerated PDHG (PA-PDHG), SPDHG, primal accelerated SPDHG (PA-SPDHG).*

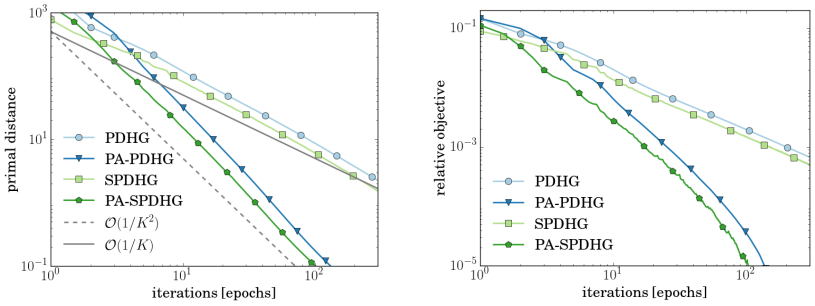

FIGURE 5. *Primal acceleration for TV denoising.* ***Left:*** *primal distance to saddle point* $\frac{1}{2}\|x^K - x^\sharp\|^2$ ***Right:*** *relative objective* $\frac{\Phi(x^K)-\Phi^\sharp}{\Phi(x^0)-\Phi^\sharp}$

$\mathbb{R}^{d_1 \times d_2}, d_1, d_2 = 612$ where the image is first convolved with a motion blur of size 9×9 and then the outer 5 pixels of each boundary are clipped. Thereby we do not assume any boundary condition on the inner part of the reconstructed image of size 602×602. The noise is modeled to be Poisson with a constant background of 30 compared to the approximate data mean of 478.96. We further assume to have the knowledge that the reconstructed image should be non-negative and upper-bounded by 100. We want to solve the following minimization problem

$$\min_{x \in \mathbb{X}} \left\{ f_1(\mathbf{A}x) + \alpha \sum_{i=1}^{2} \xi_\eta(\nabla_i x) + \imath_C(x) \right\}.$$

Here the data fidelity f_1 is the Kullback–Leibler divergence (27) where the sum is taken over all pixels. The prior information is the anisotropic TV with Huberized norm

$$\xi_\eta(y) = \sum_{i,j} \begin{cases} |y_{i,j}| & \text{if } |y_{i,j}| > \eta \\ \frac{1}{2\eta}|y_{i,j}|^2 + \frac{\eta}{2} & \text{else} \end{cases}$$

with Huber parameter $\eta = 1$, regularization parameter $\alpha = 0.1$ and constraint set $C = \{x \in \mathbb{X} : 0 \leq x_{i,j} \leq 100\}$.

By the nature of the forward operator we have that $\mathbf{A}x \geq 0$ whenever $x \geq 0$. Therefore the solution to the optimization problem remains the same if we replace the Kullback–Leibler divergence by the differentiable

$$f_1(y) = \sum_{i=1}^{N} \begin{cases} y_i + r_i - b_i + b_i \log\left(\frac{b_i}{y_i + r_i}\right) & \text{if } y_i \geq 0 \\ \frac{b_i}{2r_i^2} y_i^2 + \left(1 - \frac{b_i}{r_i}\right) y_i + r_i - b_i + b_i \log\left(\frac{b_i}{r_i}\right) & \text{else} \end{cases} \tag{30}$$

which has a $\left(\max_i b_i / r_i^2\right)$ Lipschitz continuous gradient. The Lipschitz constant is well-defined and non-zero as both the data b_i as well as the background r_i are positive.

To obtain the saddle point problem we dualize data term and Huberized TV and obtain the saddle point problem over $\mathbb{X}$ and $\mathbb{Y} = \mathbb{R}^N \times \mathbb{X} \times \mathbb{X}$

$$\min_{x \in \mathbb{X}} \sup_{y \in \mathbb{Y}} \left\{ \langle y_1, \mathbf{A}x \rangle - f_1^*(y_1) + \sum_{i=2}^{3} \langle y_i, \nabla_{i-1} x \rangle - \imath_{\alpha B}(y_i) - \frac{\eta}{2\alpha} \|y_i\|_2^2 + \imath_C(x) \right\}.$$

The convex conjugate of the modified Kullback–Leibler divergence (30) is

$$f_1^*(z) = \sum_{i=1}^{N} \begin{cases} \frac{r_i^2}{2b_i} z_i^2 + \left(r_i - \frac{r_i^2}{b_i}\right) z_i + \frac{r_i^2}{2b_i} + \frac{3b_i}{2} - 2r_i - b_i \log\left(\frac{b_i}{r_i}\right) & \text{if } z_i < 1 - \frac{b_i}{r_i} \\ -r_i z_i - b_i \log(1 - z_i) & \text{if } 1 - \frac{b_i}{r_i} \leq z_i < 1 \\ \infty & \text{if } z_i \geq 1 \end{cases}$$

which can readily seen to be $\left(\min_i \frac{r_i^2}{b_i}\right)$-strongly convex. Moreover, for the algorithm we also need its proximal operator

$$\left[\operatorname{prox}_{f_1^*}^{\sigma}(z)\right]_i = \begin{cases} \frac{b_i z_i - \sigma r_i b_i + \sigma r_i^2}{b_i + \sigma r_i^2} & \text{if } z_i < 1 - \frac{b_i}{r_i} \\ \frac{1}{2}\left\{ z_i + \sigma r_i + 1 - \sqrt{(z_i + \sigma r_i - 1)^2 + 4\sigma b_i} \right\} & \text{else} \end{cases}.$$

Note that $(\alpha \xi_\eta)^*(y) = \imath_{\alpha B}(y_i) - \frac{\eta}{2\alpha} \|y_i\|_2^2$ where $\imath_{\alpha B}$ is the indicator function of the ℓ^∞ ball with radius α around 0, cf. (29).

The norms of the directional derivatives are 2 and the norm of the blurring operator was estimated to be 10.45 by the power method.

Parameters. In this experiment we choose $\gamma = 1$ and consider both uniform ($p_i = 1/n$) and importance sampling ($p_i = \|A_i\| / \sum_j \|A_j\|$). The latter will have probabilities proportional to the norm of the operators. The number of subsets are either $n = 1$ in the deterministic case or $n = 3$ in the stochastic case. The (initial) step size parameters are chosen to be

- PDHG: $\sigma = 1/\|\mathbf{A}\|$, $\tau = 1/\|\mathbf{A}\|$
- DA-PDHG: $\tilde{\sigma}^0 = \frac{\mu_f}{\|\mathbf{A}\|}$, $\tau^0 = 1/\|\mathbf{A}\|$
- SPDHG (uniform sampling): $\sigma_i = 1/\|\mathbf{A}_i\|$, $\tau = 1/(n \max_i \|\mathbf{A}_i\|)$
- SPDHG (importance): $\sigma_i = 1/\|\mathbf{A}_i\|$, $\tau = 1/\sum_i \|\mathbf{A}_i\|$
- DA-SPDHG (uniform): $\tilde{\sigma}^0 = \min_i \frac{\mu_i p_i^2}{\tau^0 \|\mathbf{A}_i\|^2 + 2\mu_i p_i (1 - p_i)}$, $\tau^0 = 1/(n \max_i \|\mathbf{A}_i\|)$
- DA-SPDHG (importance): $\tilde{\sigma}^0 = \min_i \frac{\mu_i p_i^2}{\tau^0 \|\mathbf{A}_i\|^2 + 2\mu_i p_i (1 - p_i)}$, $\tau^0 = 1/\sum_i \|\mathbf{A}_i\|$

Results. Figure 6 shows the data and reconstructed images from PDHG, PA-PDHG and DA-SPDHG with importance sampling after 100 epochs. It is easy to see that while PDHG has not restored the contrast yet and both deterministic variants are still relatively blurry, DA-SPDHG with importance sampling yields a sharp reconstruction. The quantitative results in Figure 7 confirm the visual assessment. The combination of importance sampling and acceleration yields a significant speed up in both quality measures.

FIGURE 6. *Deblurring with total variation regularization with Poisson noise and unknown boundaries. Results after 100 epochs.* ***Top:*** *From left to right: Blurry and noisy data with kernel (magnified), PDHG, DA-PDHG and DA-SPDHG (importance sampling, 3 subsets).* ***Bottom:*** *Close-ups of top row.*

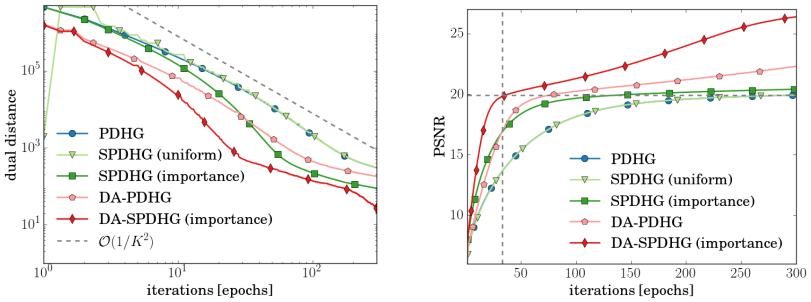

FIGURE 7. *Quantitative results for deblurring. In both quality measures (****left:*** *distance to dual part of the saddle point* $\frac{1}{2}\|y^K - y^\sharp\|^2$ *and* ***right:*** *peak signal-to-noise ratio (PSNR)) the accelerated algorithms are faster than the non-accelerated counterparts. While uniform sampling does not speed up the convergence for this example, the combination of importance sampling with acceleration yields the by far fastest algorithm.*

7.4. PET Reconstruction (Linear Rate). For the final example we turn back to PET reconstruction but this time with linear rate. This means we want to solve the same minimization problem as in the first example, but now we replace the Kullback–Leibler functional by its modified version as in the previous example. We note again that this does not change the solution of the minimization problem. Moreover, to make the TV strongly convex we add another regularization $\frac{\mu}{2}\|x\|_2^2$. Note that the proximal operator of the TV (indeed any functional) with added squared ℓ^2-norm, i.e. $g(x) = \alpha\,\mathrm{TV}(x) + \frac{\mu}{2}\|x\|_2^2$, can be solved by means of the original proximal operator $\mathrm{prox}_g^\sigma(z) = \mathrm{prox}_{\mathrm{TV}}^{\frac{\sigma\alpha}{1+\sigma\mu}}\left(\frac{z}{1+\sigma\mu}\right)$. The regularization parameters are chosen as $\alpha = 4$ and $\mu = 10$.

We will consider two different settings. First, all views of the PET forward operator are equally divided making it unnecessary to consider samplings other than uniform. Second, to test the impact of sampling, we choose one subset to contain half the number of views and divide the remaining views uniformly among the remaining subsets. This imbalanced setting should make it crucial to consider non-uniform sampling strategies.

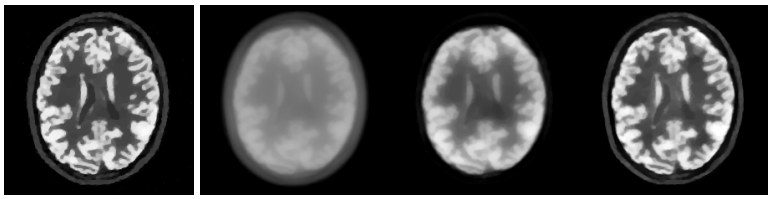

FIGURE 8. *PET reconstruction with a strongly convex TV prior and uniform sampling. Results after 5 epochs.* ***Left:*** *approximate primal part of saddle point* ***Right:*** *From left to right: PDHG, SPDHG (10 subsets) and SPDHG (250). It is clear that with more subsets, the reconstruction becomes closer to the desired solution.*

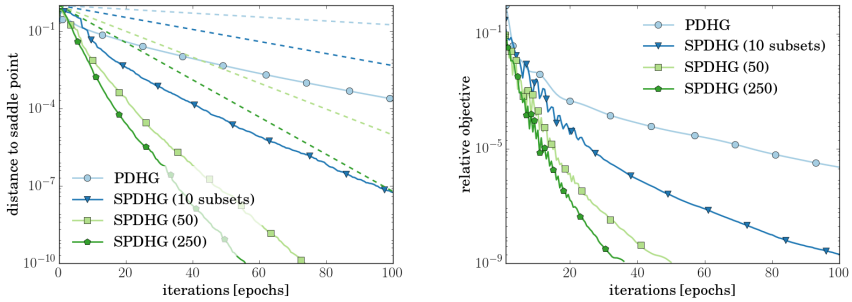

FIGURE 9. *Comparisons of PDHG and SPDHG with uniform sampling for PET reconstruction. As can be seen in terms of distance to the saddle point (**left**) and the relative objective function value (**right**) the stochastic variants are all much faster than the deterministic PDHG. Moreover, the graph on the left numerically verifies the result of Theorem 6.1 as the empirical solid curves lie all below their provable worst case upper bound (dashed).*

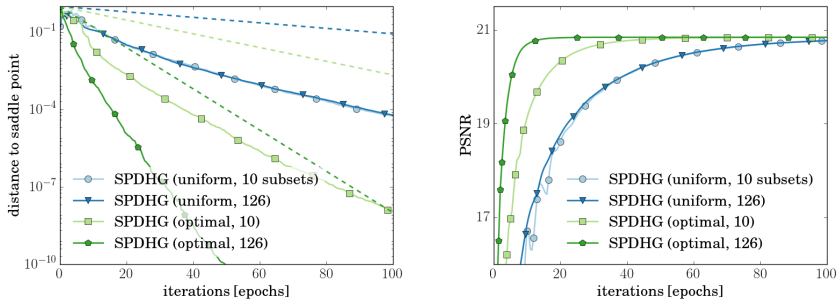

FIGURE 10. *Comparison of different samplings for SPDHG with imbalanced subsets. As it can be seen from either graph, non-uniform subset selections require a non-uniform sampling for improved convergence speeds.*

Parameters. In this experiment we choose $\rho = 0.99$ and the sampling to be either uniform or optimal. The other step size parameters are chosen as derived in subsection 6.2.

Results. The visual results in Figure 8 show that SPDHG is much faster than the deterministic PDHG and that the speed increases by increasing the number of subsets. This is quantitatively confirmed in Figure 9 in terms of distance to the saddle point and objective function value. Moreover, the impact of optimal sampling is apparent in Figure 10 which shows the performance of SPDHG in the imbalanced setting.

8. Conclusions and Future Work. We proposed a natural stochastic generalization of the deterministic PDHG algorithm to convex-concave saddle point problems that are separable in the dual variable. The analysis was carried out in the context of *arbitrary samplings* which enabled us to obtain known deterministic convergence results as special cases. We proposed optimal choices of the step size parameters with which the proposed algorithm showed superior empirical performance on a variety of optimization problems in imaging.

In the future, we would like to extend the analysis to include iteration dependent (adaptive) probabilities [17] and strong convexity parameters to further exploit the structure of many relevant problems. Moreover, the present optimal sampling strategies are only for scalar-valued step sizes and serial sampling. In the future, we wish to extend this to other sampling strategies such as multi-block or parallel sampling.

Appendix A. Auxiliary Results for the Proofs.

Proof of Lemma 4.2. Denote by $\hat{y}_i^k$ the value of y^k if it got updated (and thus is not random with respect to $\mathbb{S}^k$). Furthermore, as $y_i^k = y_i^{k-1}$ for $i \notin \mathbb{S}^k$ and $y_i^k = \hat{y}_i^k$ if $i \in \mathbb{S}^k$ we have by completing the norm for any $x \in \mathbb{X}$

$$\begin{aligned}\langle x, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle &= \langle x, \sum_{i\in\mathbb{S}^k} \mathbf{A}_i^* p_i^{-1}(\hat{y}_i^k - y_i^{k-1})\rangle \\ &= \left\langle \sqrt{c}\mathbf{T}^{-1/2}x, \sum_{i\in\mathbb{S}^k} \mathbf{C}_i^* \frac{\mathbf{S}_i^{-1/2}}{\sqrt{cp_i}}(\hat{y}_i^k - y_i^{k-1})\right\rangle \\ &\geq -\frac{c}{2}\|x\|^2_{\mathbf{T}^{-1}} - \frac{1}{2c}\left\|\sum_{i\in\mathbb{S}^k} \mathbf{C}_i^*(p_i\mathbf{S}_i)^{-1/2}(\hat{y}_i^k - y_i^{k-1})\right\|^2 .\end{aligned} \tag{31}$$

Moreover, we now estimate the expectation of the second term of the right hand side of (31)

$$\begin{aligned}\mathbb{E}^k\left\|\sum_{i\in\mathbb{S}^k} \mathbf{C}_i^*(p_i\mathbf{S}_i)^{-1/2}(\hat{y}_i^k - y_i^{k-1})\right\|^2 &\leq \sum_{i=1}^n p_i v_i \|(p_i\mathbf{S}_i)^{-1/2}(\hat{y}_i^k - y_i^{k-1})\|^2 \\ &\leq \gamma^2 \sum_{i=1}^n p_i \|\hat{y}_i^k - y_i^{k-1}\|^2_{(\mathbf{S}_i p_i)^{-1}} \\ &= \gamma^2 \mathbb{E}^k\|y^k - y^{k-1}\|^2_{(\mathbf{SP})^{-1}}\end{aligned} \tag{32}$$

where the first inequality is due to the ESO inequality (4), the second inequality holds by the definition of γ and the last equation holds due to Corollary A.4. Combining the expected value of inequality (31) with inequality (32) yields the assertion. ∎

LEMMA A.1 (Deterministic inequality). *Let* $x^{k+1} := \operatorname{prox}_g^{\mathbf{T}}\left(x^k - \mathbf{T}\mathbf{A}^*\overline{y}^k\right)$ *and* $\hat{y}_i^{k+1} := \operatorname{prox}_{f_i^*}^{\mathbf{S}_i}\left(y_i^k + \mathbf{S}_i\mathbf{A}_i x^{k+1}\right)$. *Moreover, let* $(x^\sharp, y^\sharp)$ *be a saddle point. Then*

$$\begin{aligned}&\frac{1}{2}\|x^k - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{\mathbf{S}^{-1}} \\ &\geq D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp) + D_{f^*}^{\mathfrak{f}^\sharp}(\hat{y}^{k+1}, y^\sharp) - \langle \mathbf{A}(x^{k+1} - x^\sharp), \hat{y}^{k+1} - \overline{y}^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathbf{T}^{-1}} \\ &\quad + \frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{\mathbf{T}^{-1}+\mathbf{M}_g} + \frac{1}{2}\|y^k - \hat{y}^{k+1}\|^2_{\mathbf{S}^{-1}} + \frac{1}{2}\|\hat{y}^{k+1} - y^\sharp\|^2_{\mathbf{S}^{-1}+\mathbf{M}_f} .\end{aligned}$$

Proof. Standard calculations for this type of algorithm lead to

$$
\begin{aligned}
g(x^\sharp) &\geq g(x^{k+1}) + \langle \mathbf{T}^{-1}(x^k - x^{k+1}) - \mathbf{A}^*\overline{y}^k, x^\sharp - x^{k+1}\rangle + \frac{\mu_g}{2}\|x^\sharp - x^{k+1}\|^2 \\
f_i^*(y_i^\sharp) &\geq f_i^*(\hat{y}_i^{k+1}) + \langle \mathbf{S}_i^{-1}(y_i^k - \hat{y}_i^{k+1}) + \mathbf{A}_i x^{k+1}, y_i^\sharp - \hat{y}_i^{k+1}\rangle + \frac{\mu_i}{2}\|y^\sharp - \hat{y}^{k+1}\|^2
\end{aligned} \tag{33}
$$

for $i = 1, \dots, n$. Summing the inequalities (33) and exploiting the identity $2\langle \mathbf{M}(a-b), c-b\rangle = \|a-b\|^2_{\mathbf{M}} + \|b-c\|^2_{\mathbf{M}} - \|a-c\|^2_{\mathbf{M}}$ yields

$$
\begin{aligned}
&\frac{1}{2}\|x^k - x^\sharp\|^2_{\mathbf{T}^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{\mathbf{S}^{-1}} \\
&\geq g(x^{k+1}) - g(x^\sharp) + \sum_{i=1}^n f_i^*(\hat{y}_i^{k+1}) - f_i^*(y_i^\sharp) + \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathbf{T}^{-1}} \\
&\quad + \frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{\mathbf{T}^{-1}+\mathbf{M}_g} + \frac{1}{2}\|y^k - \hat{y}^{k+1}\|^2_{\mathbf{S}} + \frac{1}{2}\|\hat{y}^{k+1} - y^\sharp\|^2_{\mathbf{S}^{-1}+\mathbf{M}_f} \\
&\quad + \langle \mathbf{A}x^{k+1}, y^\sharp - \hat{y}^{k+1}\rangle - \langle \mathbf{A}(x^\sharp - x^{k+1}), \overline{y}^k\rangle
\end{aligned}
$$

where we used the definition of the inner product and the norm on the product space $\mathbb{Y}$. It now suffices to complete the Bregman distances $D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp) = g(x^{k+1}) - g(x^\sharp) - \langle -\mathbf{A}^* y^\sharp, x^{k+1} - x^\sharp\rangle$ and $D_{f^*}^{\mathfrak{f}^\sharp}(\hat{y}^{k+1}, y^\sharp) = \sum_{i=1}^n f_i^*(\hat{y}_i^{k+1}) - f_i^*(y_i^\sharp) - \langle \mathbf{A}_i x^\sharp, \hat{y}_i^{k+1} - y_i^\sharp\rangle$. ∎

LEMMA A.2 (Stochastic inequality). *Let $x^{k+1}, \hat{y}^{k+1}$ be defined as in Lemma A.1, $y^{k+1}, \overline{y}^{k+1}$ as in Algorithm Algorithms 2 and 3 and $(x^\sharp, y^\sharp)$ be a saddle point. Moreover, let γ^2 bound some ESO parameters of $(\mathbf{S}^k)^{1/2}\mathbf{P}^{-1/2}\mathbf{A}(\mathbf{T}^k)^{1/2}$. Then*

$$
\begin{aligned}
&\mathbb{E}^k\mathbb{E}^{k-1}\left\{\frac{1}{2}\|x^k - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}+2\mathbf{M}_f(\mathbf{P}^{-1}-\mathbf{I})}\right. \\
&\qquad \left. - \theta^{k-1}\langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1}\rangle + \frac{(\theta^{k-1}\gamma)^2}{2}\|y^k - y^{k-1}\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}}\right\} \\
&\geq \mathbb{E}^{k+1}\mathbb{E}^k\left\{\frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}+2\mathbf{M}_g} + \frac{1}{2}\|y^{k+1} - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}+2\mathbf{M}_f\mathbf{P}^{-1}}\right. \\
&\qquad \left. - \langle x^{k+1} - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^{k+1} - y^k)\rangle + \frac{1}{2}\|y^{k+1} - y^k\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}}\right\}.
\end{aligned}
$$

Proof. We adapt the result of Lemma A.1 to the stochastic setting suitable for strongly convex functionals. In addition, all step size parameters may vary along the iterations. First, by Lemma A.1 and the strong convexity of f and g, which implies $D_g^{\mathfrak{g}^\sharp}(x^{k+1}, x^\sharp) \geq \frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{\mathbf{M}_g}, D_{f^*}^{\mathfrak{f}^\sharp}(\hat{y}^{k+1}, y^\sharp) \geq \frac{1}{2}\|\hat{y}^{k+1} - y^\sharp\|^2_{\mathbf{M}_f}$ it follows that

$$
\begin{aligned}
&\frac{1}{2}\|x^k - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{(\mathbf{S}^k)^{-1}} \\
&\geq -\langle \mathbf{A}(x^{k+1} - x^\sharp), \hat{y}^{k+1} - \overline{y}^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_{(\mathbf{T}^k)^{-1}} \\
&\quad + \frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}+2\mathbf{M}_g} + \frac{1}{2}\|y^k - \hat{y}\|^2_{(\mathbf{S}^k)^{-1}} + \frac{1}{2}\|\hat{y} - y^\sharp\|^2_{(\mathbf{S}^k)^{-1}+2\mathbf{M}_f}
\end{aligned}
$$

and invoking Corollary A.4 yields

$$\begin{aligned}
&\frac{1}{2}\|x^k - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}} + \frac{1}{2}\|y^k - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}+2\mathbf{M}_f(\mathbf{P}^{-1}-\mathbf{I})} \\
&\quad - \theta^{k-1}\langle x^k - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle \\
&\geq \mathbb{E}^{k+1}\left\{\frac{1}{2}\|x^{k+1} - x^\sharp\|^2_{(\mathbf{T}^k)^{-1}+2\mathbf{M}_g} + \frac{1}{2}\|y^{k+1} - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}+2\mathbf{M}_f\mathbf{P}^{-1}}\right. \\
&\quad - \langle x^{k+1} - x^\sharp, \mathbf{A}^*\mathbf{P}^{-1}(y^{k+1} - y^k)\rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_{(\mathbf{T}^k)^{-1}} \\
&\quad \left. + \frac{1}{2}\|y^{k+1} - y^k\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}} + \theta^{k-1}\langle x^{k+1} - x^k, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle\right\}.
\end{aligned}$$

where we used the extrapolation $\overline{y}^k = y^k + \theta^{k-1}\mathbf{P}^{-1}(y^k - y^{k-1})$ and the identity

$$\|y^k - y^\sharp\|^2_{(\mathbf{S}^k)^{-1}} + \|y^k - y^\sharp\|^2_{((\mathbf{S}^k)^{-1}+2\mathbf{M}_f)(\mathbf{P}^{-1}-\mathbf{I})} = \|y^k - y^\sharp\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}+2\mathbf{M}_f(\mathbf{P}^{-1}-\mathbf{I})}\,.$$

Taking the expectations $\mathbb{E}^k, \mathbb{E}^{k-1}$ and estimating the last inner product by Lemma 4.2 as $\theta^{k-1}\langle x^{k+1} - x^k, \mathbf{A}^*\mathbf{P}^{-1}(y^k - y^{k-1})\rangle \geq -\frac{1}{2}\|x^{k+1} - x^k\|^2_{(\mathbf{T}^k)^{-1}} - \frac{(\theta^{k-1}\gamma)^2}{2}\|y^k - y^{k-1}\|^2_{(\mathbf{S}^k\mathbf{P})^{-1}}$ yields the assertion. ∎

This next statement is trivial.

LEMMA A.3. *Let $\mathbb{S}$ be a Bernoulli random variable that is 1 with probability p and 0 with probability $1-p$. Let y^+ be a random vector that is updated to $\hat{y}$ if $\mathbb{S} = 1$ and set to y^0 if $\mathbb{S} = 0$. Then, for any $y \in \mathbb{Y}$ and any mapping φ on $\mathbb{Y}$ the equality $\mathbb{E}_{\mathbb{S}}\varphi(y^+) = p\varphi(\hat{y}) + (1-p)\varphi(y^0)$ holds.*

COROLLARY A.4. *Let $\mathbb{S} \subset \{1, \dots, n\}$ be a random set and denote the probability that an index i is in $\mathbb{S}$ by $p_i := \mathbb{P}(i \in \mathbb{S})$. Let $y^+ \in \mathbb{Y} = \prod_{i=1}^n \mathbb{Y}_i$ be a random vector of a product space such that its ith component is updated to $\hat{y}_i$ if $i \in \mathbb{S}$ and set to y_i^0 otherwise. Let $\mathbf{P} := \operatorname{diag}(p_1\mathbf{I}, \dots, p_n\mathbf{I})$ be a block diagonal operator and $\mathbf{M}$ be any other block-diagonal operator. Then for any $y \in \mathbb{Y}$ we have the following three identities:*

$$\begin{aligned}
\|\hat{y} - y\|^2_{\mathbf{M}} &= \mathbb{E}_{\mathbb{S}}\|y^+ - y\|^2_{\mathbf{M}\mathbf{P}^{-1}} - \|y^0 - y\|^2_{\mathbf{M}\mathbf{P}^{-1}(\mathbf{I}-\mathbf{P})}\,, \\
\|\hat{y} - y^0\|^2_{\mathbf{M}} &= \mathbb{E}_{\mathbb{S}}\|y^+ - y^0\|^2_{\mathbf{M}\mathbf{P}^{-1}}, \quad \text{and} \quad \hat{y} = \mathbf{P}^{-1}\mathbb{E}_{\mathbb{S}}y^+ - (\mathbf{P}^{-1} - \mathbf{I})y^0\,.
\end{aligned}$$

REFERENCES

[1] J. ADLER, H. KOHR, AND O. ÖKTEM, *Operator Discretization Library (ODL)*, 2017, https://github.com/odlgroup/odl.

[2] Z. ALLEN-ZHU, Y. YUAN, P. RICHTÁRIK, AND Y. YUAN, *Even Faster Accelerated Coordinate Descent Using Non-Uniform Sampling*, International Conference on Machine Learning, 48 (2016), https://arxiv.org/abs/1512.09103.

[3] M. S. C. ALMEIDA AND M. A. T. FIGUEIREDO, *Frame-based image deblurring with unknown boundary conditions using the alternating direction method of multipliers*, IEEE Transactions on Image Processing, 22 (2013), pp. 582–585, https://doi.org/10.1109/ICIP.2013.6738120.

[4] P. BALAMURUGAN AND F. BACH, *Stochastic Variance Reduction Methods for Saddle-Point Problems*, (2016), pp. 1–23, https://arxiv.org/abs/1605.06398.

[5] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2011, https://doi.org/10.1007/978-1-4419-9467-7.

[6] A. Beck and M. Teboulle, *A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[7] M. Benning, C.-B. Schönlieb, T. Valkonen, and V. Vlačić, *Explorations on anisotropic regularisation of dynamic inverse problems by bilevel optimisation*. 2016, https://arxiv.org/abs/1602.01278.

[8] D. P. Bertsekas, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey*, in Optimization for Machine Learning, S. Sra, S. and Nowozin, S. and Wright, ed., MIT Press, 2011, pp. 85–120.

[9] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical Programming, 129 (2011), pp. 163–195, https://doi.org/10.1007/s10107-011-0472-0.

[10] D. Blatt, A. O. Hero, and H. Gauchman, *A Convergent Incremental Gradient Method with a Constant Step Size*, SIAM Journal on Optimization, 18 (2007), pp. 29–51, https://doi.org/10.1137/040615961.

[11] K. Bredies and M. Holler, *A TGV-Based Framework for Variational Image Decompression, Zooming, and Reconstruction. Part I: Analytics*, SIAM Journal on Imaging Sciences, 8 (2015), pp. 2814–2850, https://doi.org/10.1137/15M1023865.

[12] A. Chambolle, *An Algorithm for Total Variation Minimization and Applications*, Journal of Mathematical Imaging and Vision, 20 (2004), pp. 89–97.

[13] A. Chambolle and T. Pock, *A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging*, Journal of Mathematical Imaging and Vision, 40 (2011), pp. 120–145, https://doi.org/10.1007/s10851-010-0251-1.

[14] A. Chambolle and T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319, https://doi.org/10.1017/S096249291600009X.

[15] A. Chambolle and T. Pock, *On the ergodic convergence rates of a first-order primal-dual algorithm*, vol. 159, Springer Berlin Heidelberg, 2016, https://doi.org/10.1007/s10107-015-0957-3.

[16] P. L. Combettes and J.-C. Pesquet, *Stochastic Quasi-Fejér Block-Coordinate Fixed Point Iterations with Random Sweeping*, SIAM Journal on Optimization, 25 (2015), pp. 1221–1248, https://doi.org/10.1137/140971233.

[17] D. Csiba, Z. Qu, and P. Richtárik, *Stochastic Dual Coordinate Ascent with Adaptive Probabilities*, Proceedings of The 32nd International Conference on Machine Learning, 37 (2015), pp. 674–683.

[18] C. D. Dang and G. Lan, *Randomized Methods for Saddle Point Computation*, (2014), pp. 1–29, https://arxiv.org/abs/1409.8625.

[19] R. M. de Oliviera, E. S. Helou, and E. F. Costa, *String-averaging incremental subgradients for constrained convex optimization with applications to reconstruction of tomographic images*, Inverse Problems, 32 (2016), p. 115014, https://doi.org/10.1088/0266-5611/32/11/115014.

[20] A. Defazio, F. Bach, and S. Lacoste-Julien, *SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives*, Nips, (2014), pp. 1–12, https://arxiv.org/abs/arXiv:1407.0202v2.

[21] E. Esser, X. Zhang, and T. F. Chan, *A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science*, SIAM Journal on Imaging Sciences, 3 (2010), pp. 1015–1046, https://doi.org/10.1137/09076934X.

[22] V. Estellers, S. Soatto, and X. Bresson, *Adaptive Regularization With the Structure Tensor*, IEEE Transactions on Image Processing, 24 (2015), pp. 1777–1790, https://doi.org/10.1109/TIP.2015.2409562.

[23] O. Fercoq and P. Bianchi, *A Coordinate Descent Primal-Dual Algorithm with Large Step Size and Possibly Non Separable Functions*. 2015, https://arxiv.org/abs/1508.04625.

[24] O. Fercoq and P. Richtárik, *Accelerated, Parallel and PROXimal Coordinate Descent*, SIAM Journal on Optimization, 25 (2015), pp. 1997–2023.

[25] X. Gao, Y. Xu, and S. Zhang, *Randomized Primal-Dual Proximal Block Coordinate Updates*, arXiv preprint arXiv:1605.05969, (2016), https://arxiv.org/abs/1605.05969.

[26] G. Gilboa, M. Moeller, and M. Burger, *Nonlinear Spectral Analysis via One-homogeneous Functionals - Overview and Future Prospects*. 2015, https://arxiv.org/abs/1510.01077.

[27] F. Knoll, M. Holler, T. Koesters, R. Otazo, K. Bredies, and D. K. Sodickson, *Joint MR-PET reconstruction using a multi-channel image regularizer*, IEEE Transactions on Medical Imaging, 0062, https://doi.org/10.1109/TMI.2016.2564989.

[28] Konečný, J. Liu, P. Richtárik, and M. Takáč, *Mini-batch Semi-Stochastic Gradient Descent in the Proximal Setting*, IEEE Journal of Selected Topics in Signal Processing, 10 (2016),

pp. 242–255.
[29] R. D. Kongskov, Y. Dong, and K. Knudsen, *Directional Total Generalized Variation Regularization*, 2 (2017), pp. 1–24, https://arxiv.org/abs/1701.02675.
[30] P.-L. Lions and B. Mercier, *Splitting Algorithms for the Sum of Two Nonlinear Operators*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 964–979.
[31] A. Nedić and D. P. Bertsekas, *Incremental subgradient methods for nondifferentiable optimization*, SIAM J. Optimization, 12 (2001), pp. 109–138, https://doi.org/10.1109/CDC.1999.832908.
[32] J. M. Ollinger and J. A. Fessler, *Positron Emisson Tomography*, IEEE Signal Processing Magazine, 14 (1997), pp. 43–55, https://doi.org/10.1109/79.560323.
[33] N. Parikh and S. P. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization, 1 (2014), pp. 123–231, https://doi.org/10.1561/2400000003.
[34] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin, *Coordinate Friendly Structures, Algorithms and Applications*, Annals of Mathematical Sciences and Applications, 1 (2016), pp. 1–54, https://doi.org/10.4310/AMSA.2016.v1.n1.a2.
[35] J.-C. Pesquet and A. Repetti, *A Class of Randomized Primal-Dual Algorithms for Distributed Optimization*, (2015), https://arxiv.org/abs/1406.6404.
[36] T. Pock and A. Chambolle, *Diagonal preconditioning for first order primal-dual algorithms in convex optimization*, Proceedings of the IEEE International Conference on Computer Vision, (2011), pp. 1762–1769, https://doi.org/10.1109/ICCV.2011.6126441.
[37] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the Mumford-Shah functional*, Proceedings of the IEEE International Conference on Computer Vision, (2009), pp. 1133–1140, https://doi.org/10.1109/ICCV.2009.5459348.
[38] Z. Qu and P. Richtárik, *Coordinate Descent with Arbitrary Sampling I: Algorithms and Complexity*, Optimization Methods and Software, (2014), p. 32, https://doi.org/10.1080/10556788.2016.1190360.
[39] Z. Qu and P. Richtárik, *Quartz: Randomized Dual Coordinate Ascent with Arbitrary Sampling*, Neural Information Processing Systems, (2015), pp. 1–34.
[40] P. Richtárik and M. Takáč, *Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function*, Mathematical Programming, 144 (2014), pp. 1–38, https://doi.org/10.1007/s10107-012-0614-z.
[41] P. Richtárik and M. Takáč, *On optimal probabilities in stochastic coordinate descent methods*, Optimization Letters, 10 (2016), pp. 1233–1243, https://doi.org/10.1007/s11590-015-0916-1.
[42] P. Richtárik and M. Takáč, *Parallel coordinate descent methods for big data optimization*, Mathematical Programming, 156 (2016), pp. 433–484.
[43] D. Rigie and P. La Riviere, *Joint Reconstruction of Multi-Channel, Spectral CT Data via Constrained Total Nuclear Variation Minimization*, Physics in Medicine and Biology, 60 (2015), pp. 1741–1762, https://doi.org/10.1088/0031-9155/60/4/1741.
[44] L. Rosasco and S. Villa, *Stochastic Inertial primal-dual algorithms*, (2015), pp. 1–15, https://arxiv.org/abs/arXiv:1507.00852v1.
[45] M. Schmidt, N. Le Roux, and F. Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming, (2016), pp. 1–30, https://doi.org/10.1007/s10107-016-1030-6.
[46] M. Takáč, A. Bijral, P. Richtárik, and N. Srebro, *Mini-batch Primal and Dual Methods for SVMs*, in In Proceedings of the 30th International Conference on Machine Learning, 2013.
[47] P. Tseng, *An Incremental Gradient(-Projection) Method with Momentum Term and Adaptive Stepsize Rule*, SIAM Journal on Optimization, 8 (1998), pp. 506–531.
[48] T. Valkonen, *Block-proximal methods with spatially adapted acceleration*, (2016), https://arxiv.org/abs/1609.07373.
[49] M. Wen, S. Yue, Y. Tan, and J. Peng, *A randomized intertial primal-dual fixed point algorithm for monotone inclusions*, (2016), pp. 1–26, https://arxiv.org/abs/1611.05142.
[50] Y. Zhang and L. Xiao, *Stochastic Primal-Dual Coordinate Method for Regularized Empirical Risk Minimization*, Proceedings of the 32nd International Conference on Machine Learning - ICML '15, (2015), pp. 1–34.
[51] L. W. Zhong and J. T. Kwok, *Fast Stochastic Alternating Direction Method of Multipliers*, Journal of Machine Learning Research, 32 (2014), pp. 46—-54.
[52] Z. Zhu and A. J. Storkey, *Adaptive Stochastic Primal-Dual Coordinate Descent for Separable Saddle Point Problems*, in Machine Learning and Knowledge Discovery in Databases, A. Appice, P. P. Rodrigues, V. S. Costa, C. Soares, J. Gama, and A. Jorge, eds., Porto, 2015, Springer, pp. 643–657.