Citation for published version:
Gutiérrez, EB, Delplancke, C & Ehrhardt, MJ 2021, Convergence Properties of a Randomized Primal-Dual Algorithm with Applications to Parallel MRI. in A Elmoataz, J Fadili, Y Quéau, J Rabin & L Simon (eds), *Scale Space and Variational Methods in Computer Vision - 8th International Conference, SSVM 2021, Proceedings.* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12679 LNCS, Springer Science and Business Media Deutschland GmbH, pp. 254-266, 8th International Conference on Scale Space and Variational Methods in Computer Vision, SSVM 2021, Virtual, Online, 16/05/21. https://doi.org/10.1007/978-3-030-75549-2_21

DOI:
10.1007/978-3-030-75549-2_21

Publication date:
2021

Document Version
Peer reviewed version

This is a post-peer-review, pre-copyedit version of an article published in SSVM 2021: Scale Space and Variational Methods in Computer Vision. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-75549-2_21

ON THE CONVERGENCE OF THE STOCHASTIC PRIMAL-DUAL HYBRID GRADIENT FOR CONVEX OPTIMIZATION

ERIC B. GUTIERREZ*, CLAIRE DELPLANCKE†, AND MATTHIAS J. EHRHARDT‡

Abstract. Stochastic Primal-Dual Hybrid Gradient (SPDHG) was proposed by Chambolle et al. (2018) and is a practical tool to solve nonsmooth large-scale optimization problems. In this paper we prove its almost sure convergence for convex but not necessarily strongly convex functionals. The proof makes use of a classical supermartingale result, and also rewrites the algorithm as a sequence of random continuous operators in the primal-dual space. We compare our analysis with a similar argument by Alacaoglu et al., and give sufficient conditions for an unproven claim in their proof.

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

1. Introduction.

Optimization problems have numerous applications among many fields such as imaging, data science or machine learning, to name a few. Optimization problems in data science are often formulated as

$$\hat{x} \in \arg\min_{x\in X} \sum_{i=1}^{n} f_i(A_i x) + g(x) \tag{1.1}$$

where $f_i : Y_i \to \mathbb{R}\cup\{\infty\}$ and $g : X \to \mathbb{R}\cup\{\infty\}$ are convex functionals, and $A_i : X \to Y_i$ are linear operators between finite-dimensional Hilbert spaces.

Examples of problems in this form are total variation regularized image reconstruction [13, 20] such as image denoising [8] or PET reconstruction [6]; regularized empirical risk minimization [21, 22] such as support vector machine (SVM) [3] or least absolute shrinkage and selection operator (LASSO) [5]; and optimization with large number of constraints [12, 16], among others.

Primal-dual methods offer an important advantage over gradient descent when solving this kind of problems. While some classical approaches such as gradient descent are not applicable when the functionals f_i or g are not smooth [8], primal-dual methods are able to find solutions to (1.1) without assuming differentiability. For convex, proper and lower-semicontinuous functionals f_i, g, a primal-dual formulation for (1.1) reads

$$\hat{x}, \hat{y} \in \arg\min_{x\in X}\max_{y\in Y} \sum_{i=1}^{n} \langle A_i x, y_i\rangle - f_i^*(y_i) + g(x) \tag{1.2}$$

where f^* is the *convex conjugate* of f [2], and $Y = \Pi_{i=1}^{n} Y_i$. We refer to any solution $\hat{w} = (\hat{x}, \hat{y})$ of (1.2) as a saddle point.

A well-known example of primal-dual methods that solve (1.2) is the Primal-Dual Hybrid Gradient (PDHG) [7, 11, 18], as presented by Chambolle & Pock (2011). It naturally breaks down the complexity of (1.1) into separate optimization problems by doing separate updates for the primal and dual variables x, y, as shown in (2.1). In general, PDHG is proven to converge to a solution of (1.2), however its iterations become very costly for large-scale problems, e.g. when $n \gg 1$ [6].

*Department of Mathematical Sciences, University of Bath (ebgc20@bath.ac.uk)
†Department of Mathematical Sciences, University of Bath (cd902@bath.ac.uk)
‡Institute for Mathematical Innovation. Department of Mathematical Sciences, University of Bath (M.Ehrhardt@bath.ac.uk)

More recently, Chambolle et al. proposed the Stochastic Primal-Dual Hybrid Gradient (SPDHG) [6] which reduces the per-iteration computational cost of PDHG by randomly sampling the dual variable: at each step, instead of the full dual variable y, only a random subset of its coordinates y_i gets updated. This offers significantly better performance than the deterministic PDHG for large-scale problems [6]. Examples of similar random primal-dual algorithms are found in [10, 13, 14, 15, 22].

We are interested in the convergence of SPDHG. In [6], it is shown that, for arbitrary convex functionals f_i and g, SPDHG converges in the sense of *Bregman distances*, which in general does not imply almost sure convergence in the norm. In this paper we prove the almost sure convergence of SPDHG for convex but not necessarily strongly convex functionals. The main result is stated in Section 3. A sketch of the proof is laid out in Section 4, and the complete proof is detailed in Section 5.

Additionally, in Section 6 we compare our analysis to similar arguments in the literature. In particular, we look into a claim proposed by Alacaouglu et al. [1] and offer sufficient conditions for the validity of their results.

2. The Algorithm. In order to solve (1.2), the deterministic PDHG method with dual extrapolation reads

$$\begin{aligned} x^{k+1} &= \operatorname{prox}_{\tau g}(x^k - \tau A^T \bar{y}^k) \\ y^{k+1} &= \operatorname{prox}_{\sigma f^*}(y^k + \sigma A x^{k+1}) \end{aligned} \tag{2.1}$$

where $\bar{y}^k = 2y^k - y^{k-1}$ is an extrapolation on the previous iterates, and the *proximity operator* of a convex functional f is given by

$$\operatorname{prox}_{\sigma f}(v) := \arg\min_{y \in Y} \frac{\|v-y\|^2}{2\sigma} + f(y).$$

SPDHG, in contrast, reduces the cost of iterations by only partially updating the dual variable $y = (y_i)_{i=1}^n$: at every iteration k, choose $j \in \{1, ..., n\}$ at random with probability $p_i = \mathbb{P}(j = i) > 0$, so that only the variable y_j^{k+1} is updated, while the rest remain unchanged, i.e. $y_i^{k+1} = y_i^k$ for $i \neq j$. SPDHG can be thus summarized in Algorithm 2.1.

Algorithm 2.1 SPDHG

Choose $\tau, \sigma_i > 0$ and $x^0 \in X$. Set $y^0 = \mathbf{0} \in Y$ and $z^0 = \bar{z}^0 = \mathbf{0} \in X$.
for $k \geq 0$ **do**

Select $j^k \in \{1, \dots, n\}$ at random

$$x^{k+1} = \operatorname{prox}_{\tau g}(x^k - \tau \bar{z}^k)$$

$$y_i^{k+1} = \begin{cases} \operatorname{prox}_{\sigma_i f_i^*}(y_i^k + \sigma_i A_i x^{k+1}) & \text{if } i = j^k \\ y_i^k & \text{else} \end{cases}$$

$$\delta^k = A_{j^k}^T (y_{j^k}^{k+1} - y_{j^k}^k)$$

$$z^{k+1} = z^k + \delta^k$$

$$\bar{z}^{k+1} = z^{k+1} + p_{j^k}^{-1} \delta^k$$

end for

Remark 2.1. In order to compute δ^k for the dual extrapolation $\bar{z}^{k+1}$, it is necessary to recall $y^k_{j^k}$ from memory. Only the two latest versions of each y_i need to be stored. The variables δ^k, z^k and $\bar{z}^k$ each require the same memory as x^k.

3. Main Result. We establish the almost sure convergence of SPDHG for any convex functionals, under the same step size conditions as in [6]:

Assumption 3.1. We assume the following to hold:
1. The functionals g, f_i are convex, proper and lower-semicontinuous.
2. The step sizes $\tau, \sigma_i > 0$ satisfy

$$\tau\sigma_i\|A_i\|^2 < p_i \quad \text{for every } i. \tag{3.1}$$

3. The set of solutions to (1.2) is nonempty.

THEOREM 3.2 (**Convergence of SPDHG**). *Let $(w^k)_{k\in\mathbb{N}} = (x^k, y^k)_{k\in\mathbb{N}}$ be a random sequence in $\mathbb{R}^d$ generated by Algorithm 2.1. Under Assumption 3.1, the sequence $(w^k)_{k\in\mathbb{N}}$ converges almost surely to a solution of (1.2).*

4. Sketch of the Proof. The following results lay out the proof of Theorem 3.2. The complete proof is detailed in Section 5. We use the notation $\|x\|^2_T = \langle Tx, x\rangle$, as well as the block diagonal operators $Q, S : Y \to Y$ given by $Q = \operatorname{diag}(p_1^{-1}, ..., p_n^{-1})$ and $S = \operatorname{diag}(\sigma_1, ..., \sigma_n)$. The conditional expectation at time $k+1$ is denoted, for any functional φ, by

$$\mathbb{E}^{k+1}(\varphi(w^{k+1})) = \mathbb{E}(\varphi(w^{k+1})|w^k).$$

The proof of Theorem 3.2 uses the following important inequality from SPDHG, which is a consequence of ([6], Lemma 4.4). This inequality is best summarized in ([1], Lemma 4.1), which we have further simplified by using the fact that Bregman distances of convex functionals are nonnegative ([6], Section 4).

LEMMA 4.1 ([1], Lemma 4.1). *Let $(w^k)_{k\in\mathbb{N}}$ be a random sequence in $\mathbb{R}^d$ generated by Algorithm 2.1 under Assumption 3.1. Then for every saddle point $\hat{w}$,*

$$V^k(w^k - \hat{w}) \geq \mathbb{E}^{k+1}(V^{k+1}(w^{k+1} - \hat{w})) + V(x^{k+1} - x^k, y^k - y^{k-1}) \tag{4.1}$$

where the functionals V and V^k are given by

$$V(x, y) = \|x\|^2_{\tau^{-1}} + 2\langle QAx, y\rangle + \|y\|^2_{QS^{-1}}$$

$$V^k(x, y) = \|x\|^2_{\tau^{-1}} - 2\langle QAx, y^k - y^{k-1}\rangle + \|y^k - y^{k-1}\|^2_{QS^{-1}} + \|y\|^2_{QS^{-1}}.$$

The following result is the central argument of our proof. It makes use of inequality (4.1) and a classical result from Robbins & Siegmund (Lemma 5.2) to establish an important convergence result. Its proof is detailed in section 5.

PROPOSITION 4.2. *Let $(w^k)_{k\in\mathbb{N}}$ be a random sequence in $\mathbb{R}^d$ generated by Algorithm 2.1 under Assumption (3.1) and let $\hat{w}$ be a saddle point. Then:*

i) The sequence $(w^k)_{k\in\mathbb{N}}$ is a.s. bounded.
ii) The sequence $(V^k(w^k - \hat{w}))_{k\in\mathbb{N}}$ converges a.s.
iii) The sequence $(\|w^k - \hat{w}\|)_{k\in\mathbb{N}}$ converges a.s.
iv) If every cluster point of $(w^k)_{k\in\mathbb{N}}$ is a.s. a saddle point, the sequence $(w^k)_{k\in\mathbb{N}}$ converges a.s. to a saddle point.

Finally, we need to prove that every cluster point of $(w^k)_{k\in\mathbb{N}}$ is almost surely a solution to (1.2), as in Proposition 4.3. To show this, we have rewritten Algorithm 2.1 as a sequence of random operators. The details are explained in Lemma 5.5 in Section 5.

PROPOSITION 4.3. *Let $(w^k)_{k\in\mathbb{N}}$ be a random sequence in $\mathbb{R}^d$ generated by Algorithm 2.1 under Assumption 3.1. Then every cluster point of $(w^k)_{k\in\mathbb{N}}$ is almost surely a saddle point.*

Proof of Theorem 3.2. By Proposition 4.3, every cluster point of $(w^k)_{k\in\mathbb{N}}$ is almost surely a saddle point and, by Proposition 4.2 iv), the sequence $(w^k)_{k\in\mathbb{N}}$ converges almost surely to a saddle point. ∎

5. Proof of Convergence. This section contains detailed proofs for our two main arguments, Propositions 4.2 and 4.3. The proof of Proposition 4.2 follows a similar strategy to that of Combettes & Pesquet in ([9], Proposition 2.3), and we have divided it into three sections.

5.1. Proof of Proposition 4.2 i). To show this first part, we require the following lemma from [6]:

LEMMA 5.1 ([6], Lemma 4.2). *Let $p_i^{-1}\tau\sigma_i\|A_i\|^2 \le \gamma^2 < 1$ for every i and let y^k be defined as in Algorithm 2.1. Then for every $x \in X$,*

$$\mathbb{E}^k(V(x, y^k - y^{k-1})) \ge (1-\gamma)\mathbb{E}^k\big(\|x\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\big).$$

Proof of Proposition 4.2 i). By ([1], Lemma 4.1), for any saddle point $\hat{w}$ we have

$$\Delta^k \ge \mathbb{E}^{k+1}(\Delta^{k+1}) + V(x^{k+1} - x^k, y^k - y^{k-1}) \tag{5.1}$$

where $\Delta^k = V^k(w^k - \hat{w})$. By Lemma 5.1,

$$\mathbb{E}^k(V(x^{k+1} - x^k, y^k - y^{k-1})) \ge (1-\gamma)\mathbb{E}^k\Big\{\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\Big\}.$$

Hence, taking the full expectation in (5.1) yields

$$\mathbb{E}(\Delta^k) \ge \mathbb{E}(\Delta^{k+1}) + (1-\gamma)\mathbb{E}\big(\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\big).$$

Taking the sum from $k = 0$ to $k = N-1$ gives

$$\Delta^0 \ge \mathbb{E}(\Delta^N) + (1-\gamma)\mathbb{E}\Big\{\sum_{k=0}^{N-1}\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\Big\} \tag{5.2}$$

where $y^{-1} = y^0$. This implies $\Delta^0 \ge \mathbb{E}(\Delta^N)$ and, by Lemma 5.1 we have

$$\mathbb{E}(\Delta^N) \ge \mathbb{E}\big\{(1-\gamma)(\|x^N - \hat{x}\|^2_{\tau^{-1}} + \|y^N - y^{N-1}\|^2_{QS^{-1}}) + \|y^N - \hat{y}\|^2_{QS^{-1}}\big\}. \tag{5.3}$$

It follows that

$$\Delta^0 \ge (1-\gamma)\|x^N - \hat{x}\|^2_{\tau^{-1}} + \|y^N - \hat{y}\|^2_{QS^{-1}} \quad \text{a.s.}$$

from where it is clear that the sequence $(w^N)_{N\in\mathbb{N}}$ is bounded almost surely. ∎

5.2. Proof of Proposition 4.2 ii)-iii). As in ([9], Proposition 2.3), we use the following classical result from Robbins & Siegmund.

LEMMA 5.2 ([19], Theorem 1). *Let $\mathcal{F}_k$ be a sequence of sub-σ-algebras such that $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for every k, and let α_k, η_k be nonnegative $\mathcal{F}_k$-measurable random variables such that $\sum_{k=1}^{\infty}\eta_k < \infty$ almost surely and*

$$\mathbb{E}(\alpha_{k+1}\,|\,\mathcal{F}_k) \;\le\; \alpha_k + \eta_k \quad a.s.$$

for every k. Then α_k converges almost surely to a random variable in $[0, \infty)$.

Proof of Proposition 4.2 ii)-iii). From (5.3) we have $\mathbb{E}(\Delta^N) \geq 0$. Thus taking the limit as $N \to \infty$ in (5.2) yields

$$\mathbb{E}\Big\{\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\Big\} < \infty \tag{5.4}$$

which implies

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}} < \infty \quad \text{a.s.} \tag{5.5}$$

and, in particular,

$$\|y^k - y^{k-1}\|_{QS^{-1}} \to 0 \quad \text{a.s.} \tag{5.6}$$

Since $(w^k)_{k\in\mathbb{N}}$ is bounded a.s., so is $(x^k)_{k\in\mathbb{N}}$ and, since the operators Q, A and S are also bounded, there exists $M > 0$ such that, for every k,

$$|\langle QA(x^k - \hat{x}), y^k - y^{k-1}\rangle| \leq \|QA\|\|x^k - \hat{x}\|\|y^k - y^{k-1}\| \leq M\|y^k - y^{k-1}\|_{QS^{-1}}$$

a.s. and therefore, by (5.6),

$$\langle QA(x^k - \hat{x}), y^k - y^{k-1}\rangle \to 0 \quad \text{a.s.} \tag{5.7}$$

The fact that $(w^k)_{k\in\mathbb{N}}$ is a.s. bounded, together with (5.6) and (5.7) imply the sequence $(\Delta^k)_{k\in\mathbb{N}}$ is also a.s. bounded. Thus there exists $\hat{M} \geq 0$ such that $\Delta^k + \hat{M} \geq 0$ for every k. Let $\alpha_k = \Delta^k + \hat{M}$ and $\eta_k = 2|\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle|$. From (5.1) we deduce

$$\alpha_k + \eta_k \;\geq\; \mathbb{E}^{k+1}(\alpha_{k+1}) \quad \text{a.s. for every } k, \tag{5.8}$$

where all the terms are nonnegative and, for some $\tilde{M} > 0$,

$$\begin{aligned}\eta_k = 2|\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle| &\leq 2\|QA\|\|x^{k+1} - x^k\|\|y^k - y^{k-1}\| \\ &\leq 2\tilde{M}\|x^{k+1} - x^k\|_{\tau^{-1}}\|y^k - y^{k-1}\|_{QS^{-1}} \\ &\leq \tilde{M}\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\end{aligned}$$

which implies, by (5.5), $\sum_{k=1}^{\infty} \eta_k < \infty$ a.s.. Thus (5.8) satisfies all the assumptions of Lemma 5.2 and it yields

$$\Delta^k \to \alpha \quad \text{a.s.}$$

for some $\alpha \in [-\hat{M}, \infty)$. Furthermore, from (5.6) and (5.7) we know some of the terms in Δ^k converge to 0 almost surely, namely

$$-2\langle QA(x^k - \hat{x}), y^k - y^{k-1}\rangle + \|y^k - y^{k-1}\|^2_{QS^{-1}} \to 0 \quad \text{a.s.}$$

hence

$$\|x^k - \hat{x}\|^2_{\tau^{-1}} + \|y^k - \hat{y}\|^2_{QS^{-1}} \to \alpha \quad \text{a.s.}$$

Finally, the norm $\|w\|^2_R := \|x\|^2_{\tau^{-1}} + \|y\|^2_{QS^{-1}}$ is equivalent to the norm in $\mathbb{R}^d$. Since the sequence $(\|w^k - \hat{w}\|_R)_{k\in\mathbb{N}}$ converges almost surely, so does $(\|w^k - \hat{w}\|)_{k\in\mathbb{N}}$. ∎

5.3. Proof of Proposition 4.2 iv). The two following lemmas are consequence of ([9], Proposition 2.3) and their proofs are included in the appendix for completeness. We use the standard notation $(\mathbf{\Omega}, \mathcal{F}, P)$ for the probability space corresponding to the random iterations w^k.

LEMMA 5.3 ([9], Proposition 2.3 (iii)). *Let $\mathbf{F}$ be a closed subset of $\mathbb{R}^d$ and let $(w^k)_{k\in\mathbb{N}}$ be a sequence of random variables such that the sequence $(\|w^k - w\|)_{k\in\mathbb{N}}$ converges almost surely for every $w \in \Phi$. Then there exists $\Omega \in \mathcal{F}$ such that $\mathbb{P}(\Omega) = 1$ and the sequence $(\|w^k(\omega) - w\|)_{k\in\mathbb{N}}$ converges for all $\omega \in \Omega$ and $w \in \mathbf{F}$.*

LEMMA 5.4 ([9], Proposition 2.3 (iv)). *Let $\mathbf{G}(w^k)$ be the set of cluster points of a random sequence $(w^k)_{k\in\mathbb{N}}$ in $\mathbb{R}^d$. Assume there exists $\Omega \in \mathcal{F}$ such that $\mathbb{P}(\Omega) = 1$ and for every $\omega \in \Omega$, $\mathbf{G}(w^k(\omega))$ is nonempty and the sequence $(\|w^k(\omega) - w\|)_{k\in\mathbb{N}}$ converges for all $w \in \mathbf{G}(w^k(\omega))$. Then $(w^k)_{k\in\mathbb{N}}$ converges almost surely to an element of $\mathbf{G}(w^k)$.*

Proof of Proposition 4.2 iv). Let $\mathbf{F}$ be the set of solutions to the saddle point problem (1.2). By Proposition 4.2 (iii) and Lemma 5.3, there exists $\Omega \in \mathcal{F}$ such that the sequence $(\|w^k(\omega) - w\|)_{k\in\mathbb{N}}$ converges for every $w \in \mathbf{F}$ and $\omega \in \Omega$. This implies, since $\mathbf{F}$ is nonempty, that $(w^k(\omega))_{k\in\mathbb{N}}$ is bounded and thus $\mathbf{G}(w^k(\omega))$ is nonempty for all $\omega \in \Omega$. By assumption, there exists $\tilde{\Omega} \in \mathcal{F}$ such that $\mathbf{G}(w^k(\omega)) \subset \mathbf{F}$ for every $\omega \in \tilde{\Omega}$. Let $\omega \in \Omega \cap \tilde{\Omega}$, then $(\|w^k(\omega) - w\|)_{k\in\mathbb{N}}$ converges for every $w \in \mathbf{G}(w^k(\omega)) \neq \emptyset$. By Lemma 5.4, we get the result. ∎

5.4. Proof of Proposition 4.3. In order to prove Proposition 4.3 we require the following lemma.

LEMMA 5.5. *Denote $w = (w_i)_{i=0}^n = (x, y_1, ..., y_n)$ and for every $j \in \{1, ..., n\}$ let the operator $T_j : \mathbb{R}^d \to \mathbb{R}^d$ be defined by*

$$(T_j w)_0 = \operatorname{prox}_{\tau g}\big(x - \tau A^T y - \big(1 + \frac{1}{p_j}\big)\tau A_j^T((T_j w)_j - y_j)\big)$$

$$(T_j w)_i = \begin{cases} \operatorname{prox}_{\sigma_i f_i^*}(y_i + \sigma_i A_i x) & \text{if } i = j \\ y_i & \text{else} \end{cases} \qquad \text{for } 1 \leq i \leq n.$$

Then the iterations w^k generated by Algorithm 2.1 satisfy

$$T_{j^k}(x^{k+1}, y^k) = (x^{k+2}, y^{k+1}). \tag{5.9}$$

Furthermore, $\hat{w}$ is a solution to the saddle point problem (1.2) if and only if it is a fixed point of T_j for each $j \in \{1, ..., n\}$.

Proof. By definition of the iterates in Algorithm 2.1, $(T_{j^k}(x^{k+1}, y^k))_i = y_i^{k+1}$ for every i and we have

$$\begin{aligned}
\bar{z}^{k+1} &= z^k + (1 + \frac{1}{p_{j^k}})\delta^k \\
&= A^T y^k + (1 + \frac{1}{p_{j^k}})A_{j^k}^T(y_{j^k}^{k+1} - y_{j^k}^k) \\
&= A^T y^k + (1 + \frac{1}{p_{j^k}})A_{j^k}^T((T_{j^k}(x^{k+1}, y^k))_{j^k} - y_{j^k}^k).
\end{aligned}$$

Thus $(T_{j^k}(x^{k+1}, y^k))_0 = \operatorname{prox}_{\tau g}(x^{k+1} - \tau \bar{z}^{k+1}) = x^{k+2}$, which proves (5.9). Now let w be a fixed point of T_j for every j. Then, for any j,

$$y_j = w_j = (T_j w)_j = \operatorname{prox}_{\sigma_i f_j^*}(y_j + \sigma_j A_j x),$$

from where it follows that, for every j,

$$
\begin{aligned}
x = w_0 = (T_j w)_0 &= \operatorname{prox}_{\tau g}(x - \tau A^* y - (1 + \frac{1}{p_j})\tau A_j^*((T_j w)_j - y_i) \\
&= \operatorname{prox}_{\tau g}(x - \tau A^* y).
\end{aligned}
$$

These conditions on x and y define a saddle point ([4], 6.4.2). The converse result is direct. ∎

Proof of Proposition 4.3. Let j^k be the sampling generated by the algorithm and denote $z^k = (x^{k+1}, y^k)$. By Lemma 5.5 we have $z^{k+1} = T_{j^k} z^k$ and, by (5.4),

$$
\mathbb{E}(\|z^k - z^{k-1}\|^2) = \mathbb{E}(\|x^{k+1} - x^k\|^2 + \|y^k - y^{k-1}\|^2) \to 0. \tag{5.10}
$$

Furthermore, by the properties of the conditional expectation,

$$
\begin{aligned}
\mathbb{E}(\|z^{k+1} - z^k\|^2) = \mathbb{E}(\mathbb{E}^k(\|z^{k+1} - z^k\|^2)) &= \mathbb{E}\Big(\sum_{j=1}^n \mathbb{P}(j^k = j)\|T_j z^k - z^k\|^2\Big) \\
&= \sum_{j=1}^n \mathbb{P}(j^k = j)\mathbb{E}(\|T_j z^k - z^k\|^2).
\end{aligned}
$$

By assumption $p_j = \mathbb{P}(j^k = j) > 0$, thus by (5.10) we have $\mathbb{E}(\|T_j z^k - z^k\|^2) \to 0$ for every j and therefore

$$
T_j z^k - z^k \to 0 \quad \text{a.s. for every } j \in \{1, ..., n\}. \tag{5.11}
$$

Assume now a convergent subsequence $w^{\ell_k} \to w^*$. From (5.10), $y^k - y^{k-1} \to 0$ a.s. and so z^{ℓ_k} also converges to w^*. By (5.11) and the continuity of T_j, for every j there holds

$$
w^* = \lim_{k\to\infty} z^{\ell_k} = \lim_{k\to\infty} T_j z^{\ell_k} = T_j\big(\lim_{k\to\infty} z^{\ell_k}\big) = T_j w^* \quad \text{a.s.}
$$

Hence w^* is almost surely a fixed point of T_j for each j and, by Lemma 5.5, w^* is a saddle point. ∎

6. Relation to Other Work.

6.1. Chambolle et al. (2018). In the original paper for SPDHG [6], it is shown that, under Assumption 3.1, the *Bregman distance* to any solution $\hat{x}, \hat{y}$ of (1.2) converges to zero, i.e. the iterates x^k, y^k of Algorithm 2.1 satisfy

$$
D_g^{-A^T\hat{y}}(x^k, \hat{x}) + D_{f^*}^{A\hat{x}}(y^k, \hat{y}) \to 0 \quad \text{a.s.,} \tag{6.1}
$$

where the Bregman distance $D_h^q(u, v)$ is defined by

$$
D_h^q(u, v) = h(u) - h(v) - \langle q, u - v\rangle
$$

for any functional h and any point $q \in \partial h(v)$ in the subdifferential of h. In [6], it is shown that (6.1) implies $(x^k, y^k) \to (\hat{x}, \hat{y})$ a.s. if f_i or g are strongly convex.

Another difference with [6] is that we limit ourselves to serial sampling in Algorithm 2.1. In this context, condition (3.1) is equivalent to the step-size condition of the original SPDHG result ([6], Theorem 4.3).

6.2. Combettes & Pesquet (2014). In [9], Combettes & Pesquet look into the convergence of random sequences $(w^k)_{k\in\mathbb{N}}$ of the form

$$w_i^{k+1} = \begin{cases} (Tw^k)_i & \text{if } i \in \mathbb{S}^k \\ w_i^k & \text{else} \end{cases} \tag{6.2}$$

where $\mathbb{S}^k \subset \{1, ..., n\}$ is chosen at random. They use the Robbins-Siegmund lemma (Lemma 5.2) to prove that, for any nonexpansive operator T, the sequence $(w^k)_{k\in\mathbb{N}}$ converges to a fixed point of T. Later, Pesquet & Repetti [17] used this to prove convergence for a wide class of random algorithms of the form (6.2), where $T = (I + B)^{-1}$ is the resolvent operator of a maximally monotone operator B.

For the case of SPDHG, which involves a sequence of operators T^k, we take inspiration from ([9], Proposition 3.2), which states sufficient conditions on the sequence $(w^k)_{k\in\mathbb{N}}$ that guarantee its convergence to a fixed point $w \in \bigcap_{k\in\mathbb{N}} \mathrm{Fix} T^k$.

6.3. Alacaoglu et al. (2019). Recently Alacaoglu et al. proposed a proof for the almost sure convergence of SPDHG ([1], Theorem 4.4) using a strategy similar to ours. However, their argument is based on the following claim: Under the step size condition (3.1), the iterates of Algorithm 2.1 satisfy

$$V(x^{k+1} - x^k, y^k - y^{k-1}) \geq (1-\gamma)\left(\|x^{k+1} - x^k\|^2_{\tau^{-1}} + \|y^k - y^{k-1}\|^2_{QS^{-1}}\right) \tag{6.3}$$

where γ is such that $p_i^{-1}\tau\sigma_i\|A_i\|^2 \leq \gamma^2 < 1$, and V is defined in Lemma 4.1. No proof is offered in [1] for this claim, although it is used in their proof of the almost sure convergence of SPDHG. Here, we show via an example how inequality (6.3) does not hold for arbitrary x, y under assumption (3.1), and we propose sufficient conditions under which it does.

Lemma 6.1. *The following two assertions hold:*

1. *Let step size condition (3.1) be satisfied, then V may be negative, i.e. there exist x, y such that $V(x, y) < 0$.*
2. *Assume $\sum_{i=1}^n p_i^{-1}\tau\sigma_i\|A_i\|^2 < 1$. Then for all x, y,*

$$V(x, y) \geq (1-\gamma)(\|x\|^2_{\tau^{-1}} + \|y\|^2_{QS^{-1}}).$$

Proof. Let $\gamma_i^2 = p_i^{-1}\tau\sigma_i\|A_i\|^2$. By assumption, $\gamma_i^2 < 1$. Rewrite $V(x, y)$ as

$$V(x, y) = \|x\|^2_{\tau^{-1}} + 2\langle QAx, y\rangle + \|y\|^2_{QS^{-1}} = \|\tilde{x}\|^2 + 2\langle C\tilde{x}, \tilde{y}\rangle + \|\tilde{y}\|^2$$

for $\tilde{x} = \tau^{-1/2}x$, $\tilde{y} = Q^{1/2}S^{-1/2}y$ and $C = Q^{1/2}S^{1/2}A\tau^{1/2}$. Assume $X = Y_i$ for every i and let each $A_i = I$ be the identity. Then $C_i = p_i^{-1/2}\sigma_i^{1/2}\tau^{1/2}I$ and

$$\langle C\tilde{x}, \tilde{y}\rangle = \sum_{i=1}^n \langle C_i\tilde{x}, \tilde{y}_i\rangle = \sum_{i=1}^n \langle p_i^{-1/2}\sigma_i^{1/2}\tau^{1/2}\tilde{x}, \tilde{y}_i\rangle = \sum_{i=1}^n \gamma_i\langle\tilde{x}, \tilde{y}_i\rangle$$

Choose y such that $y_i = -p_i^{1/2}\sigma^{1/2}\tau^{-1/2}x$ for every i. Then $\tilde{y}_i = -\tilde{x}$ for every i and $\|\tilde{y}\|^2 = \sum_{i=1}^n \|\tilde{y}_i\|^2 = n\|\tilde{x}\|^2$, thus

$$\begin{aligned} 2V(x, y) &= \|\tilde{x}\|^2 + 2\langle C\tilde{x}, \tilde{y}\rangle + \|\tilde{y}\|^2 = \|\tilde{x}\|^2 + 2\sum_{i=1}^n \gamma_i\langle\tilde{x}, \tilde{y}_i\rangle + \|\tilde{y}\|^2 \\ &= \|\tilde{x}\|^2 - 2\sum_{i=1}^n \gamma_i\|\tilde{x}\|^2 + n\|\tilde{x}\|^2 = \|\tilde{x}\|^2(1 + n - 2\sum_{i=1}^n \gamma_i). \end{aligned}$$

By assumption $\gamma_i < 1$, however the last term is negative if $\sum_{i=1}^n \gamma_i \geq \frac{n+1}{2}$. E.g. for $n > 1$, taking $\tau = \sigma_i = 0.9 p_i^{1/2}$ yields $\gamma_i = 0.9 < 1$ which satisfies assumption (3.1), however $\sum_{i=1}^n \gamma_i = 0.9n > \frac{n+1}{2}$. This proves the first part of the lemma.

Now assume $\sum_{i=1}^n \gamma_i^2 < 1$. For arbitrary A_i, choosing $\|\tilde{x}\| \leq 1$ yields

$$\|C\tilde{x}\|^2 = \sum_{i=1}^n \|C_i \tilde{x}\|^2 = \sum_{i=1}^n p_i^{-1} \tau \sigma_i \|A_i \tilde{x}\|^2 \leq \sum_{i=1}^n \gamma_i^2$$

from where it follows that $\|C\|^2 \leq \sum_{i=1}^n \gamma_i^2 < 1$ and thus, for any x, y,

$$\begin{aligned} V(x,y) &= \|\tilde{x}\|^2 + 2\langle C\tilde{x}, \tilde{y}\rangle + \|\tilde{y}\|^2 \geq \|\tilde{x}\|^2 - 2\|C\|\|\tilde{x}\|\|\tilde{y}\| + \|\tilde{y}\|^2 \\ &\geq \|\tilde{x}\|^2 - \|C\|(\|\tilde{x}\|^2 + \|\tilde{y}\|^2) + \|\tilde{y}\|^2 \\ &= (1 - \|C\|)\big(\|\tilde{x}\|^2 + \|\tilde{y}\|^2\big) = (1 - \|C\|)\big(\|x\|^2_{\tau^{-1}} + \|y\|^2_{QS^{-1}}\big). \end{aligned}$$

∎

Appendix A. Postponed Proofs.

Proof of Lemma 5.3. Let Z be a countable set such that $\bar{Z} = \Phi$. By assumption, for every $z \in \Phi$ there exists a set $\Omega_z \in \mathcal{F}$ such that $\mathbb{P}(\Omega_z) = 1$ and

$$\|w^k(\omega) - z\| \to \tau_z(\omega) \quad \forall \omega \in \Omega_z$$

for some random variable $\tau_z : \boldsymbol{\Omega} \to [0, \infty)$. Let $\Omega = \bigcap_{z \in Z} \Omega_z$ and let Ω^c be its complement. Then, since Z is countable,

$$\mathbb{P}(\Omega) = 1 - \mathbb{P}(\Omega^c) = 1 - \mathbb{P}\Big(\bigcup_{z \in Z} \Omega_z^c\Big) \geq 1 - \sum_{z \in Z} \mathbb{P}(\Omega_z^c) = 1.$$

Now let $z \in \Phi$ be fixed. By density there exists a sequence $(z^n)_{n \in \mathbb{N}}$ in Z such that $z^n \to z$. As just seen, for every $n \in \mathbb{N}$ there exists $\tau_n : \boldsymbol{\Omega} \to [0, \infty)$ such that $\|w^k(\omega) - z^n\| \to \tau_n(\omega)$ for every $\omega \in \Omega_{z^n}$. Let $\omega \in \Omega$, then for every $k, n \in \mathbb{N}$,

$$-\|z^n - z\| \leq \|w^k(\omega) - z\| - \|w^k(\omega) - z^n\| \leq \|z^n - z\|.$$

Hence for every $n \in \mathbb{N}$ we have

$$\begin{aligned} -\|z^n - z\| &\leq \liminf_{k \to \infty} \|w^k(\omega) - z\| - \tau_n(\omega) \\ &\leq \limsup_{k \to \infty} \|w^k(\omega) - z\| - \tau_n(\omega) \leq \|z^n - z\|. \end{aligned}$$

Taking the limit as $n \to \infty$ yields

$$\liminf_{k \to \infty} \|w^k(\omega) - z\| = \limsup_{k \to \infty} \|w^k(\omega) - z\| = \lim_{n \to \infty} \tau_n(\omega).$$

We conclude that $(\|w^k(\omega) - z\|)_{k \in \mathbb{N}}$ is a convergent sequence for every $\omega \in \Omega$. ∎

Proof of Lemma 5.4. Let $\omega \in \Omega$. From the assumptions it follows that the sequence $(w^k(\omega))_{k \in \mathbb{N}}$ is bounded, so it suffices to show that it has at most one cluster point ([2], Lemma 2.46).

Let z_1, z_2 be two cluster points of the sequence $(w^k(\omega))_{k\in\mathbb{N}}$. By assumption the sequences $(\|w^k(\omega) - z_1\|)_{k\in\mathbb{N}}$ and $(\|w^k(\omega) - z_2\|)_{k\in\mathbb{N}}$ converge, and thus the fact that

$$2\langle w^k(\omega), z_1 - z_2\rangle = \|w^k(\omega) - z_2\|^2 - \|w^k(\omega) - z_1\|^2 + \|z_1\|^2 - \|z_2\|^2$$

implies the sequence $\langle w^k(\omega), z_1 - z_2\rangle_{k\in\mathbb{N}}$ also converges to some ϱ. However, by definition of z_1 there exists a subsequence $(w^{\ell_k}(\omega))_{k\in\mathbb{N}}$ which converges to z_1, thus $\langle z_1, z_1 - z_2\rangle = \varrho$. By the same argument, $\langle z_2, z_1 - z_2\rangle = \varrho$ and we have

$$0 = \langle z_1 - z_2, z_1 - z_2\rangle = \|z_1 - z_2\|^2,$$

i.e. $z_1 = z_2$. ∎

REFERENCES

[1] A. Alacaoglu, O. Fercoq, and V. Cevher, *On the convergence of stochastic primal-dual hybrid gradient*, arXiv preprint arXiv:1911.00799, (2019).

[2] H. H. Bauschke, P. L. Combettes, et al., *Convex analysis and monotone operator theory in Hilbert spaces*, vol. 408, Springer, 2011.

[3] B. E. Boser, I. M. Guyon, and V. N. Vapnik, *A training algorithm for optimal margin classifiers*, in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[4] K. Bredies and D. Lorenz, *Mathematical Image Processing*, Springer, 2018.

[5] V. Cevher, S. Becker, and M. Schmidt, *Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics*, IEEE Signal Processing Magazine, 31 (2014), pp. 32–43.

[6] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, *Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications*, SIAM Journal on Optimization, 28 (2018), pp. 2783–2808.

[7] 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.

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

[9] 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.

[10] C. Dang and G. Lan, *Randomized first-order methods for saddle point optimization*, arXiv preprint arXiv:1409.8625, (2014).

[11] 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.

[12] O. Fercoq, A. Alacaoglu, I. Necoara, and V. Cevher, *Almost surely constrained convex optimization*, arXiv preprint arXiv:1902.00126, (2019).

[13] O. Fercoq and P. Bianchi, *A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions*, SIAM Journal on Optimization, 29 (2019), pp. 100–134.

[14] X. Gao, Y.-Y. Xu, and S.-Z. Zhang, *Randomized primal–dual proximal block coordinate updates*, Journal of the Operations Research Society of China, 7 (2019), pp. 205–250.

[15] P. Latafat, N. M. Freris, and P. Patrinos, *A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization*, IEEE Transactions on Automatic Control, 64 (2019), pp. 4050–4065.

[16] A. Patrascu and I. Necoara, *Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization*, The Journal of Machine Learning Research, 18 (2017), pp. 7204–7245.

[17] J.-C. Pesquet and A. Repetti, *A class of randomized primal-dual algorithms for distributed optimization*, arXiv preprint arXiv:1406.6404, (2014).

[18] T. Pock, D. Cremers, H. Bischof, and A. Chambolle, *An algorithm for minimizing the mumford-shah functional*, in 2009 IEEE 12th International Conference on Computer Vision, 2009, pp. 1133–1140.

[19] H. Robbins and D. Siegmund, *A convergence theorem for non negative almost supermartingales and some applications*, in Optimizing methods in statistics, Elsevier, 1971, pp. 233–257.

[20] L. I. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Physica D: nonlinear phenomena, 60 (1992), pp. 259–268.

[21] S. Shalev-Shwartz and T. Zhang, *Stochastic dual coordinate ascent methods for regularized loss minimization*, Journal of Machine Learning Research, 14 (2013), pp. 567–599.

[22] Y. Zhang and L. Xiao, *Stochastic primal-dual coordinate method for regularized empirical risk minimization*, The Journal of Machine Learning Research, 18 (2017), pp. 2939–2980.