On the convergence and sampling of randomized primal-dual algorithms and their application to parallel MRI reconstruction

Eric B Gutierrez, Claire Delplancke and Matthias J Ehrhardt

Department of Mathematical Sciences, University of Bath, BA2 7AY, UK

E-mail: ebgc20@bath.ac.uk, cd902@bath.ac.uk, M.Ehrhardt@bath.ac.uk

June 2022

Abstract. The Stochastic Primal-Dual Hybrid Gradient or SPDHG is an algorithm proposed by Chambolle et al. to efficiently solve a wide class of nonsmooth large-scale optimization problems. In this paper we contribute to its theoretical foundations and prove its almost sure convergence for convex but neither necessarily strongly convex nor smooth functionals, defined on Hilbert spaces of arbitrary dimension. We also prove its convergence for any arbitrary sampling, and for some specific samplings we propose theoretically optimal step size parameters which yield faster convergence. In addition, we propose using SPDHG for parallel Magnetic Resonance Imaging reconstruction, where data from different coils are randomly selected at each iteration. We apply SPDHG using a wide range of random sampling methods. We compare its performance across a range of settings, including mini-batch size, step size parameters, and both convex and strongly convex objective functionals. We show that the sampling can significantly affect the convergence speed of SPDHG. We conclude that for many cases an optimal sampling method can be identified.

Keywords: Inverse Problems, Parallel MRI, Primal-dual, Stochastic, Optimization

1. Introduction

Inverse problems can be solved using variational regularization, generally presented as an optimization problem. In fields such as imaging, data science or machine learning, optimization challenges are generally formulated as the convex minimization problem

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

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 and bounded operators between real Hilbert spaces.

Many active areas of research exists within this framework, with problems such as regularized risk minimization [5, 7, 33, 35], heavily-constrained optimization [14, 24] or

total variation regularization image reconstruction [31]. Within the context of image reconstruction, examples of problems in the form of (1) are image denoising [10], PET reconstruction [12] and, most relevant to the present paper, parallel Magnetic Resonance Imaging (MRI) reconstruction [16, 19, 27].

Problem (1) makes no assumptions on the differentiability of the convex functionals f_i, g. In general, if these functionals are not smooth then many classical approaches such as gradient descent are not applicable [10]. In contrast, primal-dual methods are able to solve (1) even for non-smooth functionals. When f_i, g are convex, proper and lower-semicontinuous, the *saddle point problem* for (1) is

$$\hat{x}, \hat{y} \in \arg\min_{x \in X} \max_{y \in Y} \sum_{i=1}^{n} \langle A_i x, y_i \rangle - f_i^*(y_i) + g(x) \tag{2}$$

where f_i^* is the *convex conjugate* of f_i and $Y = \Pi_{i=1}^n Y_i$. We refer to a solution of (2) as a *saddle point*.

A well-known primal-dual method is the Primal-Dual Hybrid Gradient algorithm (PDHG) [9, 13, 25]. While PDHG is proven to converge to a saddle point, its iterations can be costly for large-scale problems, e.g. when $n \gg 1$ [8]. A random primal-dual method, the Stochastic Primal-Dual Hybrid Gradient algorithm (SPDHG) was proposed recently by Chambolle et al. [8]. Its main difference over PDHG is that it reduces the per-iteration computational cost by randomly sampling the dual variable, i.e. only a random subset of the dual variable gets updated at every iteration. In [8], it is shown that SPDHG offers significantly better performance than the deterministic PDHG for large-scale problems [12, 32]. Examples of other random primal-dual algorithms can be found in [15, 17, 21, 35].

In the original paper [8], SPDHG's almost sure convergence is proven for strongly convex functionals f_i^*,g. Later, for the special case of serial sampling, Alacaoglu et al. proved SPDHG's almost sure convergence for arbitrary convex functionals in finite-dimensional Hilbert spaces [2]. An alternative proof can also be found in [18].

In this paper we complete the gap on the convergence theory of SPDHG and prove its almost sure convergence for convex functionals defined on general (possibly infinite-dimensional) Hilbert spaces for any sampling. Furthermore, as a novel application of SPDHG we perform parallel MRI reconstruction on real MRI data. We show how different samplings can lead to different optimal step size conditions, and discuss how to identify an optimal sampling. For more examples on stochastic methods applied to MRI reconstruction, see [23].

2. Stochastic Primal-Dual Hybrid Gradient

SPDHG is the result of randomizing the deterministic PDHG algorithm. PDHG with dual extrapolation [9, 26] is given by

$$\begin{aligned} x^{k+1} &= \text{prox}_{\tau g}(x^k - \tau A^* \bar{y}^k) \\ y_i^{k+1} &= \text{prox}_{\sigma_i f_i^*}(y_i^k + \sigma_i A_i x^{k+1}) \qquad \text{for } i \in \{1, ..., n\} \end{aligned} \tag{3}$$

$$\bar{y}^{k+1} = y^{k+1} + \theta(y^{k+1} - y^k)$$

where $\tau, \sigma_i > 0$ are step-size parameters, $\bar{y}^{k+1}$ is an extrapolation with parameter $\theta \in [0, 1]$, operator $A^* : Y \to X$ is $A^*y = \sum_i A_i^* y_i$, and the *proximity operator* [3] is

$$\text{prox}_h(v) := \arg\min_u \frac{\|v - u\|^2}{2} + h(u).$$

This method is proven to converge for $\theta = 1$ and step size condition $\tau\sigma_i\|A\|^2 < 1$ for every $i \in \{1, ..., n\}$, for any convex functionals g, f_i^* [9].

SPDHG, in contrast, reduces the cost of (3) by updating only a random subset of the coordinates $(y_i)_{i=1}^n$. This means, at every iteration k, a subset $\mathbb{S}^k \subset \{1, ..., n\}$ is chosen at random and only the variables y_i^{k+1} for $i \in \mathbb{S}^k$ are updated, while the rest remain unchanged:

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

We assume the random variables $\mathbb{S}^k$ are independent and identically distributed. Furthermore, the sampling must be proper, i.e. it must satisfy

$$p_i := \mathbb{P}(i \in \mathbb{S}^k) > 0 \quad \text{for every } i \in \{1, ..., n\}. \tag{4}$$

The complete SPDHG algorithm is given in Algorithm 1.

Algorithm 1 SPDHG

Choose $x^0 \in X$ and $y^0 \in Y$.
Set $z^0 = \bar{z}^0 = A^*y^0$.
for $k \geq 0$ **do**
  select $\mathbb{S}^k \subset \{1, \dots, n\}$ at random
  $x^{k+1} = \text{prox}_{\tau g}(x^k - \tau\bar{z}^k)$
  $y_i^{k+1} = \begin{cases} \text{prox}_{\sigma_i f_i^*}(y_i^k + \sigma_i A_i x^{k+1}) & \text{if } i \in \mathbb{S}^k \\ y_i^k & \text{else} \end{cases}$
  $\delta_i = A_i^*(y_i^{k+1} - y_i^k)$ for all $i \in \mathbb{S}^k$
  $z^{k+1} = z^k + \sum_{i \in \mathbb{S}^k} \delta_i$
  $\bar{z}^{k+1} = z^{k+1} + \theta \sum_{i \in \mathbb{S}^k} p_i^{-1}\delta_i$
end for

In order to minimize the number of linear operations, the auxiliary variable z^k stores the current value of A^*y^k, so that only the operators A_i^* for $i \in \mathbb{S}^k$ need to be evaluated at each iteration. Similarly, $\bar{z}^{k+1}$ represents the extrapolation

$$\bar{z}^{k+1} = A^*y^{k+1} + A^*Q(y^{k+1} - y^k)$$

where $Q : Y \to Y$ is the operator defined by $(Qy)_i = p_i^{-1}y_i$. We will often refer to the fact that Q is symmetric and positive definite.

For serial sampling, where only one coordinate is selected at every iteration, i.e. $|\mathbb{S}^k| = 1$ for every k, SPDHG takes the form of Algorithm 2, which is the special case

of Algorithm 1 where we set $\mathbb{S}^k = \{j^k\}$ and only $y_{j^k}^{k+1}$ and $\delta = \delta_{j^k}$ are activated at each iteration k.

Algorithm 2 SPDHG for serial sampling

Choose $x^0 \in X$ and $y^0 \in Y$.
Set $z^0 = \bar{z}^0 = A^* y^0$.
for $k \geq 0$ **do**
  select $j^k \in \{1, \dots, n\}$ at random
  $x^{k+1} = \text{prox}_{\tau g}(x^k - \tau \bar{z}^k)$
  $$y_i^{k+1} = \begin{cases} \text{prox}_{\sigma_i f_i^*}(y_i^k + \sigma_i A_i x^{k+1}) & \text{if } i = j^k \\ y_i^k & \text{else} \end{cases}$$
  $\delta = A_{j^k}^*(y_{j^k}^{k+1} - y_{j^k}^k)$
  $z^{k+1} = z^k + \delta$
  $\bar{z}^{k+1} = z^{k+1} + \theta p_{j^k}^{-1} \delta$
end for

3. Convergence analysis

In this section we prove the almost sure convergence of SPDHG to a saddle point of the primal-dual problem (2), as well as the conditions under which this is true. Optimal step size parameters for SPDHG are explained in Section 4. Different applications of the algorithm are illustrated in Section 5.

Assumption 1. *We assume the following to hold:*

(i) The Hilbert spaces X, Y are real and separable.

(ii) The set of solutions to (2) is nonempty.

(iii) The functionals g, f_i are convex, proper and lower-semicontinuous.

(iv) The proximity operators $\text{prox}_{\tau g}$, $\text{prox}_{\sigma_i f_i^*}$ *are weakly sequentially continuous.*

Assumptions (i), (ii) and (iii) are as in the original SPDHG result ([8], Theorem 4.3), and Assumption (ii) is further discussed in [2]. Assumption (iv) is always satisfied in finite dimensions, since the proximity operator prox_h is continuous for any convex, proper and lower-semicontinuous functional h ([3], Proposition 12.28). In general, a function T is *weakly sequentially continuous* if, for every sequence $(v^k)_{k\in\mathbb{N}}$ converging weakly to v, the sequence $(T(v^k))_{k\in\mathbb{N}}$ converges weakly to $T(v)$. For examples that satisfy (iv), see ([3], Section 24).

In order to state the step size conditions, we use the following notation. For any subset $\mathcal{S} \subset \{1, \dots, n\}$, let $A_\mathcal{S} : X \to Y$ be given by $(A_\mathcal{S} x)_i = A_i x$ if $i \in \mathcal{S}$ and $(A_\mathcal{S} x)_i = 0 \in Y_i$ if $i \notin \mathcal{S}$, with adjoint $A_\mathcal{S}^* y = \sum_{i\in\mathcal{S}} A_i^* y_i$.

Similarly, denote $C_i = \tau^{1/2}\sigma_i^{1/2}A_i$ for $i \in \{1,..,n\}$, with adjoints $C_i^* = \tau^{1/2}\sigma_i^{1/2}A_i^*$, and $C_{\mathcal{S}}$ and $C_{\mathcal{S}}^*$ defined accordingly as above. For any random sampling $\mathbb{S}$, we define the step size operator D as

$$D := Q\mathbb{E}\left(C_{\mathbb{S}}C_{\mathbb{S}}^*\right)Q. \tag{5}$$

Theorem 1 (**Convergence of SPDHG**). *Let Assumption 1 be satisfied, $\theta = 1$ and*

$$\|D\| < 1. \tag{6}$$

Then Algorithm 1 converges weakly almost surely to a solution of (2).

3.1. Step size condition

In the original paper of SPDHG [8], the authors describe the step size condition using *ESO parameters* [29]. While these conditions are equivalent, we prefer to use (6) because it offers a practical way to check the validity of the step-size parameters τ, σ_i by finding the eigenvalues of D. Moreover, from [29] we know $D : Y \to Y$ can be expressed as

$$D = \begin{pmatrix} D_{11} & \cdots & D_{1n} \\ \vdots & \ddots & \vdots \\ D_{n1} & \cdots & D_{nn} \end{pmatrix}, \qquad D_{ij} = \frac{p_{ij}}{p_i p_j} C_i C_j^* \tag{7}$$

where $D_{ij} : Y_j \to Y_i$ and $p_{ij} = \mathbb{P}\left(i \in \mathbb{S}, j \in \mathbb{S}\right)$. In contrast, ESO parameters are parameters $v_1, ..., v_n$ such that

$$\mathbb{E}\left\|C_{\mathbb{S}}^* z\right\|^2 \leq \sum_{i=1}^{n} p_i v_i \|z_i\|^2 \tag{8}$$

and the step size condition for SPDHG in [8] is

$$v_i < p_i \quad \text{for all } i \in \{1, ..., n\}. \tag{9}$$

The following lemma shows this condition is equivalent to (6).

Lemma 1. *Let D be defined as in (5). Then $\|D\| < 1$ if and only if there exist ESO parameters v_i such that $v_i < p_i$ for $i \in \{1, ..., n\}$.*

Proof. Let $\|D\| < 1$. Then

$$\mathbb{E}\|C_{\mathbb{S}}^* z\|^2 = \langle z, \mathbb{E}(C_{\mathbb{S}}C_{\mathbb{S}}^*)z\rangle = \langle Q^{-1}z, DQ^{-1}z\rangle \leq \|D\|\|Q^{-1}z\|^2 = \|D\|\sum_{i=1}^{n} p_i^2\|z_i\|^2.$$

hence (8) and (9) are satisfied by choosing $v_i = \|D\|p_i$. Conversely, let v_i satisfy (8) and (9),

$$\langle Dz, z\rangle = \langle \mathbb{E}\left(C_{\mathbb{S}}C_{\mathbb{S}}^*\right)Qz, Qz\rangle = \mathbb{E}\|C_{\mathbb{S}}^* Qz\|^2 \leq \sum_{i=1}^{n} p_i v_i \|p_i^{-1}z_i\|^2 < \sum_{i=1}^{n} p_i^2\|p_i^{-1}z_i\|^2 = \|z\|^2$$

which proves $\|D\| < 1$. □

Clearly, the step size parameters τ, σ_i that satisfy (6) are not unique. In particular, if we assume the step size parameters to be uniform, i.e. $\sigma_i = \sigma$ for all i, then we write $D = \tau\sigma Q\mathbb{E}(A_{\mathbb{S}}A_{\mathbb{S}}^*)Q$ and thus it suffices to choose τ, σ such that

$$\tau\sigma\|Q\mathbb{E}(A_{\mathbb{S}}A_{\mathbb{S}}^*)Q\| < 1. \tag{10}$$

In Section 4, we will see examples on how to find optimal step sizes that comply with this condition for specific types of random samplings.

3.2. Proof of Theorem 1

The following propositions lay out the proof of Theorem 1. For brevity, we write *a.s.* instead of *almost surely*, and use the notation $w = (x, y)$.

Proposition 1. *Let $\theta = 1$ and $(w^k)_{k\in\mathbb{N}}$ a random sequence generated by Algorithm 1 under Assumption 1 and step size condition (6). The following assertions hold:*

i) The sequence $(w^{k+1} - w^k)_{k\in\mathbb{N}}$ converges a.s. to zero.

ii) The sequence $(\|w^k - \hat{w}\|)_{k\in\mathbb{N}}$ converges a.s. for every saddle point $\hat{w}$.

iii) If every weak cluster point of $(w^k)_{k\in\mathbb{N}}$ is a.s. a saddle point, the sequence $(w^k)_{k\in\mathbb{N}}$ converges weakly a.s. to a saddle point.

The key idea of the proof consists in rewriting Algorithm 1 as a sequence of operators $T_{\mathcal{S}}$ which depend on the random subsets $\mathcal{S}$. To show this, denote

$$w = (w_0, w_1, ..., w_n) = (x, y_1, ..., y_n)$$

and, for every $\mathcal{S} \subset \{1, ..., n\}$, let the operator $T_{\mathcal{S}} : X \times Y \to X \times Y$ be defined by

$$(T_{\mathcal{S}}w)_i = \begin{cases} \operatorname{prox}_{\tau g}(x - \tau A^* y - \sum_{j\in\mathbb{S}} \left(1 + p_j^{-1}\right) \tau A_j^*((T_{\mathcal{S}}w)_j - y_j)) & \text{if } i = 0 \\ \operatorname{prox}_{\sigma_i f_i^*}(y_i + \sigma_i A_i x) & \text{if } i \in \mathcal{S} \\ y_i & \text{else} \end{cases}$$

Proposition 2. *Let $\theta = 1$ and $(w^k)_{k\in\mathbb{N}}$ a random sequence generated by Algorithm 1 under Assumption 1 and step size condition (6). The following assertions hold:*

i) The iterates $(w^k)_{k\in\mathbb{N}}$ satisfy, for every $k \geq 0$,

$$T_{\mathbb{S}^k}(x^{k+1}, y^k) = (x^{k+2}, y^{k+1}). \tag{11}$$

ii) A point $w \in X \times Y$ is a saddle point if and only if it is a fixed point of $T_{\mathcal{S}}$ for every instance of $\mathbb{S}^k$.

iii) Every weak cluster point of $(w^k)_{k\in\mathbb{N}}$ is a.s. a saddle point.

Proof of Theorem 1. By Proposition 2-iii), every weak cluster point of $(w^k)_{k\in\mathbb{N}}$ is almost surely a saddle point and, by Proposition 1-iii), the sequence $(w^k)_{k\in\mathbb{N}}$ converges weakly almost surely to a saddle point. □

3.3. Proof of Propositions 1 & 2

We use the notation $\|x\|_T^2 = \langle Tx, x\rangle$ for any operator T. We also denote

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

with $S = \mathrm{diag}(\sigma_1, ..., \sigma_n)$ and A given by $(Ax)_i = A_i x$. For any $w = (x, y)$ we write

$$\Delta_{\hat{w}}^k := V(x^k - x, y^{k-1} - y^k) + \|y^k - y\|_{QS^{-1}}^2$$

and the conditional expectation at time k is denoted as $\mathbb{E}^k(w) = \mathbb{E}(w \,|\, w^0, ..., w^{k-1})$. With this notation we will make use of ([2], Lemma 4.1), which was originally obtained as a consequence of ([8], Lemma 4.4).

Lemma 2 ([2], Lemma 4.1)**.** *Let $(w^k)_{k\in\mathbb{N}}$ be a random sequence generated by Algorithm 1 under Assumption 1. Then for every saddle point $\hat{w}$,*

$$\Delta_{\hat{w}}^k \geq \mathbb{E}^{k+1}(\Delta_{\hat{w}}^{k+1}) + V(x^{k+1} - x^k, y^k - y^{k-1}). \tag{12}$$

In (12) we have simplified the original lemma by using the fact that Bregman distances of convex functionals are nonnegative [8]. We will also require the following lemma, which is equivalent to ([8], Lemma 4.2) with $\rho^2 = \max_i v_i/p_i$ and $c = \rho^{-1}$, and where we have replaced the ESO step size condition (9) with the equivalent condition (6).

Lemma 3 ([8], Lemma 4.2)**.** *Let D be defined as in (5) such that $\|D\| = \rho^2$, and let $(y^k)_{k\in\mathbb{N}}$ be obtained through Algorithm 1. Then for all $x \in X$, $k \in \mathbb{N}$,*

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

Proof of Proposition 1-i). Taking the expectation and applying Lemma 3 to (12) yields

$$\mathbb{E}(\Delta_{\hat{w}}^k) \geq \mathbb{E}(\Delta_{\hat{w}}^{k+1}) + (1-\rho)\mathbb{E}(\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2).$$

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

$$\Delta_{\hat{w}}^0 \geq \mathbb{E}(\Delta_{\hat{w}}^N) + (1-\rho)\sum_{k=0}^{N-1} \mathbb{E}\Big\{\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2\Big\} \tag{13}$$

where we define $y^{-1} = y^0$. From Lemma 3 we know $\mathbb{E}(\Delta_{\hat{w}}^N) \geq 0$, hence taking the limit as $N \to \infty$ in (13) yields $\sum_{k=0}^{\infty} \mathbb{E}(\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2) < \infty$. By the monotone convergence theorem, this is equivalent to

$$\mathbb{E}\Big\{\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2\Big\} < \infty, \tag{14}$$

which implies

$$\sum_{k=0}^{\infty} \|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2 < \infty \quad \text{a.s.} \tag{15}$$

and it follows that a.s.

$$\|x^{k+1} - x^k\|_{\tau^{-1}} \to 0 \quad \text{and} \quad \|y^k - y^{k-1}\|_{QS^{-1}} \to 0. \tag{16}$$

Since these norms are equivalent to the usual norms in X and Y, we deduce the sequence $(w^{k+1} - w^k)_{k\in\mathbb{N}}$ converges to zero almost surely. □

For the next part we require a slight variation of Lemma 3.

Lemma 4. *Let $\varphi = \max_i \|q_i^{1/2} C_i^*\|$ and let $(y^k)_{k\in\mathbb{N}}$ be the iterates defined by Algorithm 1. For all $c > 0$, $x \in X$ and $k \in \mathbb{N}$, there holds*

$$V(x, y^k - y^{k-1}) \geq (1 - \frac{n\varphi}{c})\|x\|_{\tau^{-1}}^2 + (1 - \varphi c)\|y^k - y^{k-1}\|_{QS^{-1}}^2.$$

Proof.

$$\begin{aligned}
\langle QAx, y^k - y^{k-1}\rangle &= \sum_{i\in\mathbb{S}^k} \langle q_i A_i x, y_i^k - y_i^{k-1}\rangle \\
&= \sum_{i\in\mathbb{S}^k} \langle q_i^{1/2} C_i^* \tau^{-1/2} x, q^{1/2}\sigma_i^{-1/2}(y_i^k - y_i^{k-1})\rangle \\
&\leq \sum_{i\in\mathbb{S}^k} \|q_i^{1/2} C_i^*\| \|x\|_{\tau^{-1}} \|y_i^k - y_i^{k-1}\|_{q_i\sigma_i^{-1}} \\
&\leq \max_i \|q_i^{1/2} C_i^*\| \sum_{i\in\mathbb{S}^k} \frac{1}{2}\Big(\frac{1}{c}\|x\|_{\tau^{-1}}^2 + c\|y_i^k - y_i^{k-1}\|_{q_i\sigma_i^{-1}}^2\Big) \\
&\leq \max_i \|q_i^{1/2} C_i^*\| \frac{1}{2}\Big(\frac{n}{c}\|x\|_{\tau^{-1}}^2 + c\|y^k - y^{k-1}\|_{QS^{-1}}^2\Big).
\end{aligned}$$

□

We also recall a classical result by Robbins & Siegmund.

Lemma 5 ([30], Theorem 1)**.** *Let $\mathcal{F}_k$ be a sequence of sub-σ-algebras such that, for every k, $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ and α_k, η_k are nonnegative $\mathcal{F}_k$-measurable random variables such that $\sum_{k=1}^\infty \eta_k < \infty$ a.s. and*

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

Then $(\alpha_k)_{k\in\mathbb{N}}$ converges almost surely to a random variable in $[0, \infty)$.

Proof of Proposition 1-ii). Let $c = (n+1)\varphi$. By Lemma 4,

$$\Delta_{\hat{w}}^k \geq \frac{1}{n+1}\|x^k - \hat{x}\|_{\tau^{-1}}^2 + (1 - (n+1)\varphi^2)\|y^k - y^{k-1}\|_{QS^{-1}}^2 + \|y^k - \hat{y}\|_{QS^{-1}}^2. \tag{17}$$

Since $y^k - y^{k-1} \to 0$ a.s. this implies $\Delta_{\hat{w}}^k$ is a.s. bounded from below, i.e. there exists a random variable $M_1 \geq 0$ (independent of k) such that $\alpha^k := \Delta_{\hat{w}}^k + M_1 \geq 0$ a.s. for every k. Let $\eta^k := 2|\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle|$. Then by (12),

$$\alpha^k + \eta^k \;\geq\; \mathbb{E}^{k+1}(\alpha^{k+1}) \quad \text{a.s. for every } k \tag{18}$$

where all terms are nonnegative and, for some $M_2 \geq 0$,

$$\begin{aligned}
\eta^k = 2|\langle QA(x^{k+1} - x^k), y^k - y^{k-1}\rangle| &\leq 2\|QA\|\|x^{k+1} - x^k\|\|y^k - y^{k-1}\| \\
&\leq 2M_2\|x^{k+1} - x^k\|_{\tau^{-1}}\|y^k - y^{k-1}\|_{QS^{-1}} \\
&\leq M_2(\|x^{k+1} - x^k\|_{\tau^{-1}}^2 + \|y^k - y^{k-1}\|_{QS^{-1}}^2)
\end{aligned}$$

which implies $\sum_{k=1}^\infty \eta^k < \infty$ a.s. by (15). Thus α^k satisfies Lemma 5 and we have

$$\Delta_{\hat{w}}^k \to \Delta_{\hat{w}} \quad \text{a.s. for some } \Delta_{\hat{w}} \in [-M_1, \infty).$$

Now since $(\Delta_{\hat{w}}^k)_{k\in\mathbb{N}}$ converges a.s. and is a.s. bounded, so is the right hand side of (17), and since $y^k - y^{k-1} \to 0$ a.s. we deduce that $(x^k, y^k)_{k\in\mathbb{N}}$ is a.s. bounded. Since x^k is a.s. bounded and $y^k - y^{k-1} \to 0$ a.s. it follows that

$$\langle QA(x^k - \hat{x}), y^k - y^{k-1}\rangle \to 0 \tag{19}$$

almost surely. From (16) and (19) we know some of the terms in $\Delta_{\hat{w}}^k$ converge a.s. to zero, hence the sequence $(\|x^k - \hat{x}\|_{\tau^{-1}}^2 + \|y^k - \hat{y}\|_{QS^{-1}}^2)_{k\in\mathbb{N}}$ converges a.s. to $\Delta_{\hat{w}}$. Finally, the norm $|||w|||^2 := \|x\|_{\tau^{-1}}^2 + \|y\|_{QS^{-1}}^2$ is equivalent to the product norm $\|\cdot\|^2$ in $X \times Y$. Since $(|||w^k - \hat{w}|||)_{k\in\mathbb{N}}$ converges a.s., so does $(\|w^k - \hat{w}\|)_{k\in\mathbb{N}}$. □

Proof of Proposition 1-iii). We follow closely the proof of ([11], Proposition 2.3). Let $(\mathbf{\Omega}, \mathcal{F}, \mathbb{P})$ denote the probability space corresponding to the random sequence $(w^k)_{k\in\mathbb{N}}$, let $\mathbf{F}$ be the set of solutions to (2) and let $\mathbf{G}(w^k)$ be the set of weak cluster points of the sequence $(w^k)_{k\in\mathbb{N}}$.

By Proposition 1-ii), the sequence $(\|w^k - w\|)_{k\in\mathbb{N}}$ converges a.s. for every solution $w \in \mathbf{F}$. Using ([11], Proposition 2.3-iii)), 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 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 $\mathbb{P}(\tilde{\Omega}) = 1$ and $\mathbf{G}(w^k(\omega)) \subset \mathbf{F}$ for every $\omega \in \tilde{\Omega}$. Let $\omega \in \Omega \cap \tilde{\Omega}$, then $(\|w^k(\omega) - w\|)_{k\in\mathbb{N}}$ converges for all $w \in \mathbf{G}(w^k(\omega)) \subset \mathbf{F}$. By ([11], Proposition 2.3-iv)), we have that $(w^k)_{k\in\mathbb{N}}$ converges weakly almost surely to an element of $\mathbf{G}(w^k) \subset \mathbf{F}$. □

Proof of Proposition 2-i). By definition of the iterates in Algorithm 1, we have

$$(T_{\mathbb{S}^k}(x^{k+1}, y^k))_i = y_i^{k+1} \quad \text{for every} \quad i \in \{1, ..., n\}$$

and by induction it is easy to check that $z^k = A^* y^k$ for every $k \geq 0$, hence

$$\begin{aligned}\bar{z}^{k+1} &= z^{k+1} + \sum_{i\in\mathbb{S}^k} p_i^{-1}\delta_i = z^k + \sum_{i\in\mathbb{S}^k}(1 + p_i^{-1})\delta_i = A^* y^k + \sum_{i\in\mathbb{S}^k}(1 + p_i^{-1})A_i^*(y_i^{k+1} - y_i^k) \\ &= A^* y^k + \sum_{i\in\mathbb{S}^k}(1 + p_i^{-1})A_i^*((T_{\mathbb{S}^k}(x^{k+1}, y^k))_i - y_i^k).\end{aligned}$$

Thus $(T_{\mathbb{S}^k}(x^{k+1}, y^k))_0 = \mathrm{prox}_{\tau g}(x^{k+1} - \tau\bar{z}^{k+1}) = x^{k+2}$, which proves (11). □

Proof of Proposition 2-ii). Let w be a fixed point of $T_{\mathcal{S}}$ for every instance of $\mathbb{S}^k$. By (4), for every $i \in \{1, ..., n\}$ there exists an instance $\mathcal{S}$ such that $i \in \mathcal{S}$ and

$$y_i = w_i = (T_{\mathcal{S}} w)_i = \mathrm{prox}_{\sigma_i f_i^*}(y_i + \sigma_i A_i x).$$

Furthermore, for any $\mathcal{S}$,

$$\begin{aligned}x = w_0 = (T_{\mathcal{S}} w)_0 &= \mathrm{prox}_{\tau g}(x - \tau A^* y - \sum_{i\in\mathcal{S}}(1 + \frac{1}{p_i})\tau A_i^*((T_{\mathcal{S}} w)_i - y_i)) \\ &= \mathrm{prox}_{\tau g}(x - \tau A^* y).\end{aligned}$$

These conditions on x and y define a saddle point ([6], 6.4.2). The converse result is direct. □

Proof of Proposition 2-iii). Let $v^k = (x^{k+1}, y^k)$ so that, by Proposition 2-i), we have $v^{k+1} = T_{\mathbb{S}^k} v^k$. From Proposition 1-i) we know

$$\|v^k - v^{k-1}\|^2 = \|x^{k+1} - x^k\|^2 + \|y^k - y^{k-1}\|^2 \to 0. \tag{20}$$

Taking the conditional expectation and denoting $p_{\mathcal{S}} := \mathbb{P}(\mathbb{S}^k = \mathcal{S})$ gives

$$\mathbb{E}^{k+1}\|v^{k+1} - v^k\|^2 = \mathbb{E}^{k+1}\|T_{\mathbb{S}^k} v^k - v^k\|^2 = \sum_{\mathcal{S}} p_{\mathcal{S}} \|T_{\mathcal{S}} v^k - v^k\|^2.$$

Thus by (20)

$$T_{\mathcal{S}} v^k - v^k \to 0 \quad \text{a.s.} \tag{21}$$

for every $\mathcal{S}$ such that $p_{\mathcal{S}} > 0$. Now, let $w^{\ell_k} \rightharpoonup w^*$ be a weakly convergent subsequence. From (20) we know $y^k - y^{k-1} \to 0$ a.s. and so v^{ℓ_k} also converges weakly to w^*. This, together with (21) implies, for all $u \in X \times Y$ and all $\mathcal{S}$ such that $p_{\mathcal{S}} > 0$,

$$\langle w^*, u\rangle = \lim_{k\to\infty} \langle v^{\ell_k}, u\rangle = \lim_{k\to\infty} \langle T_{\mathcal{S}} v^{\ell_k}, u\rangle \text{ a.s.}$$

and, by the weak sequential continuity of $T_{\mathcal{S}}$ (Assumption 1), it follows that

$$\lim_{k\to\infty} \langle T_{\mathcal{S}} v^{\ell_k}, u\rangle = \langle T_{\mathcal{S}}(\lim_{k\to\infty} v^{\ell_k}), u\rangle = \langle T_{\mathcal{S}} w^*, u\rangle$$

almost surely. Hence w^* is almost surely a fixed point of $T_{\mathcal{S}}$ for each instance of $\mathbb{S}^k$. By Proposition 2-ii), w^* is a saddle point. □

4. Step size parameters

Theorem 1 requires a choice of step size parameters τ, σ_i that satisfy condition (6), i.e. $\|D\| < 1$. In this section we illustrate how to choose adequate step sizes for specific examples of random samplings, starting with the general case of not necessarily strongly convex functions. Furthermore, if the functionals g, f_i^* are strongly convex, then optimal step sizes can be determined that offer linear convergence rate [8].

4.1. The general convex case

4.1.1. Serial sampling For serial sampling, a valid step size choice is given by

$$\tau\sigma_i \|A_i\|^2 < p_i \quad \text{for every } i \in \{1, ..., n\}. \tag{22}$$

Indeed this implies (6) since, for serial sampling, (22) gives

$$\|Dz\| = \mathbb{E}\|C_{\mathbb{S}}^* Q z\|^2 = \sum_{i=1}^n p_i \|C_i^* q_i z_i\|^2 = \sum_{i=1}^n q_i \tau\sigma_i \|A_i^* z_i\|^2 < \|z\|^2.$$

This insight can be generalized to mini-batch sampling as follows.

4.1.2. b-serial sampling Consider a partition of the set $\{1, ..., n\}$ into m blocks I_j, i.e.

$$\bigcup_{j=1}^{m} I_j = \{1, ..., n\} \quad \text{and} \quad I_j \cap I_l = \emptyset \quad \text{for all} \quad j \neq l.$$

At every iteration we select a single block $j \in \{1, ..., m\}$ and update every index $i \in I_j$. For such a sampling, step size condition (22) reads

$$\tau\sigma_i \|\tilde{A}_j\|^2 < \tilde{p}_j \quad \text{for } i \in I_j,\ j \in \{1, ..., m\} \tag{23}$$

where $\tilde{p}_j$ is the probability of choosing block I_j and $\tilde{A}_j : X \to \prod_{i \in I_j} Y_i$ is the operator

$$\tilde{A}_j x = (A_i x)_{i \in I_j}.$$

Notice we assumed every index i within a block I_j to have the same dual step size σ_i.

We refer to this random process as *b-serial sampling* when all subsets have size b, i.e. $|I_j| = b$ for all $j \in \{1, ..., m\}$, with $1 \leq b \leq n$.

4.1.3. b-nice sampling Consider now the random sampling $\mathbb{S}$ that randomly selects b elements from $\{1, ..., n\}$ at each iteration, i.e. every instance of $\mathbb{S}$ is a random subset of size b. Clearly at each iteration there are $\binom{n}{b}$ possible choices. This process is different from b-serial sampling, since the subsets are not disjoint and do not form a partition of $\{1, ..., n\}$.

In this paper we will assume uniform b-nice sampling, i.e. all possible instances of $\mathbb{S}$ have equal probability. In this case, it is easy to see the probabilities $p_i = \mathbb{P}(i \in \mathbb{S})$ are given by $p_i = b/n$, and condition (10) reads

$$\frac{\tau\sigma n^2}{b^2} \|\mathbb{E}(A_{\mathbb{S}} A_{\mathbb{S}}^*)\| < 1. \tag{24}$$

4.2. The strongly convex case

Algorithm 1 has linear convergence for strongly convex functionals g, f_i^* [8]. Here, we have replaced the ESO parameter condition $v_i < \theta^{-1} p_i$ from the original theorem with the equivalent condition $\|D\| < \theta^{-1}$, as explained in Lemma 1.

Lemma 6 ([8], Theorem 6.1)**.** *Let g, f_i^* be strongly convex with parameters $\mu_g, \mu_i > 0$ for $i \in \{1, ..., n\}$ and let $(\hat{x}, \hat{y})$ be the unique solution of (2). Let D satisfy*

$$\|D\| < \frac{1}{\theta} \tag{25}$$

where the extrapolation $\theta \in (0, 1)$ satisfies the lower bounds

$$\theta \geq \frac{1}{1 + 2\mu_g \tau}, \qquad \theta \geq \max_i 1 - 2\frac{\mu_i \sigma_i p_i}{1 + 2\mu_i \sigma_i}. \tag{26}$$

Then Algorithm 1 converges with rate $\mathcal{O}(\theta^k)$.

Furthermore, it is possible to estimate the optimal (smallest) value of θ by equating the lower bounds in (26) together with step size condition (25) [8]. For instance, from (10) we know that, for uniform step sizes, condition (25) is satisfied by

$$\tau\sigma\|B\|\theta = \rho^2$$

where $B = Q(\mathbb{E}(A_{\mathbb{S}}A_{\mathbb{S}}^*)Q$ and $\rho \in (0,1)$. This together with (26) yields

$$\theta = \max_i 1 - \frac{2p_i}{1+\sqrt{\beta_i}} \tag{27}$$

where $\beta_i = 1 + \|B\|p_i/(\mu_g\mu_i\rho^2)$, and the corresponding optimal step sizes are

$$\tau = \min_i \frac{\mu_g^{-1}p_i}{1-2p_i+\sqrt{\beta_i}}, \qquad \sigma = \min_i \frac{\mu_i^{-1}}{\sqrt{\beta_i}-1}. \tag{28}$$

4.2.1. Serial sampling Better theoretical convergence rates can be computed if more is known about the sampling. In particular, optimal convergence rates for uniform and non-uniform serial sampling have been proposed in [8]. These are, for uniform serial sampling,

$$\theta_{us} = 1 - \frac{2}{n + n\max_i \sqrt{\alpha_i}}$$

where $\alpha_i = 1 + \|A_i\|^2/(\mu_g\mu_i\rho^2)$, $p_i = b/n$ and step size parameters given by

$$\tau = \frac{\mu_g^{-1}}{n-2+n\max_j\sqrt{\alpha_j}}, \qquad \sigma_i = \frac{\mu_i^{-1}}{\max_j\sqrt{\alpha_j}-1}, \qquad i \in \{1,...,n\}. \tag{29}$$

and for serial sampling with optimized probabilities,

$$\theta_{os} = 1 - \frac{2}{n + \sum_{i=1}^n \sqrt{\alpha_i}}$$

where the optimal probabilities are found to be

$$p_i = \frac{1+\sqrt{\alpha_i}}{n+\sum_{j=1}^n\sqrt{\alpha_j}}, \qquad i \in \{1,...,n\}$$

and the step size parameters are

$$\tau = \frac{\mu_g^{-1}}{n-2+\sum_{j=1}^n\sqrt{\alpha_j}}, \qquad \sigma_i = \frac{\mu_i^{-1}}{\sqrt{\alpha_i}-1}, \qquad i \in \{1,...,n\}. \tag{30}$$

In general $\theta_{os} \leq \theta_{us}$, as the former imposes less restrictions over probabilities p_i.

4.2.2. b-serial sampling These last results are also useful for b-serial sampling. To see this we define $\tilde{A}_j$ and $\tilde{p}_j$ as in (23) and $\tilde{f}_j^*$ as

$$\tilde{f}_j^*(\tilde{y}_j) = \sum_{i\in I_j} f_i(y_i), \quad \tilde{y}_j \in \prod_{i\in I_j} Y_i, \quad j \in \{1,...,m\}.$$

With this notation, our original (strongly convex) saddle point problem (2) becomes

$$\hat{x},\hat{y} \in \arg\min_{x\in X}\max_{\tilde{y}\in Y}\sum_{j=1}^m \langle \tilde{A}_j x, \tilde{y}_j\rangle - \tilde{f}_j^*(\tilde{y}_i) + g(x)$$

where $\tilde{y} = (\tilde{y}_1, ..., \tilde{y}_m)$ and $\tilde{f}_j^*$ are strongly convex with parameters $\tilde{\mu}_j = \min\{\mu_i \,|\, i \in I_j\}$. Thus Lemma 6 guarantees the linear convergence of Algorithm 1 and, as before, the optimal convergence rates are

$$\theta_{us} = 1 - \frac{2}{m + m\max_j \sqrt{\tilde{\alpha}_j}}, \qquad \theta_{os} = 1 - \frac{2}{m + \sum_{j=1}^m \sqrt{\tilde{\alpha}_j}} \tag{31}$$

with $\tilde{\alpha}_j = 1 + \|\tilde{A}_j\|^2/(\mu_g \tilde{\mu}_j \rho^2)$. In both cases, the optimal step size parameters $\tau, \tilde{\sigma}_j$ are also given as in (29) or (30).

Notice how the optimal convergence rates θ_{unif} and θ_{opt} depend on b. In particular, choosing $b = n$ gives us the optimal step sizes for the PDHG algorithm (2), i.e. SPDHG with full sampling. In this case, $m = 1$ and the optimal convergence rate is

$$\theta_{us} = \theta_{os} = 1 - \frac{2}{1 + \sqrt{1 + \frac{\|A\|^2}{\mu_g \mu_{f^*} \rho^2}}} \tag{32}$$

where $\mu_{f^*} = \min\{\mu_1, ..., \mu_n\}$ is the convexity parameter of $f^*(y) = \sum_{i=1}^n f_i^*(y_i)$.

Notice as well that there is more than one way to partition the set $\{1, ..., n\}$ into m subsets I_j of size b. For instance, if n is divisible by b, the number $\mathbf{k}$ of different partitions of $\{1, ..., n\}$ into subsets of size b is

$$\mathbf{k}(n, b) = \prod_{j=1}^{\frac{n}{b}} \begin{pmatrix} jb - 1 \\ b - 1 \end{pmatrix}. \tag{33}$$

Moreover, the optimal convergence rates θ_{us} and θ_{os} will also depend on which partition we use, since the subsets I_j define in turn the values $\tilde{\mu}_j$ and $\|\tilde{A}_j\|$. In Section 5, we will see examples of how using a different partition of $\{1, ..., n\}$ can improve the convergence rate of SPDHG with b-serial sampling.

4.2.3. b-nice sampling For uniform b-nice sampling we use rate (27) with uniform probability $p_i = b/n$, and we denote

$$\theta_{un} = 1 - \frac{2b}{n + n\max_i \sqrt{\beta_i}} \tag{34}$$

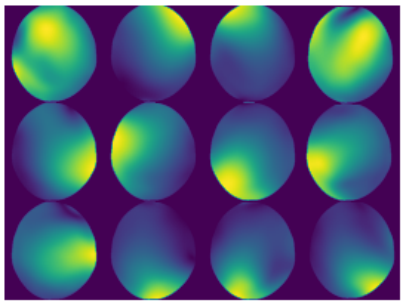
with step size parameters given as in (28).

This quantity also depends strongly on the batch size b since the probabilities p_i and thus also the norm of the operator B are determined by b. In particular, choosing $b = n$ in (34) results in $p_i = 1$ and $\|B\| = \|AA^*\| = \|A\|^2$, hence we recover the same full sampling convergence rate (32) from b-serial sampling.

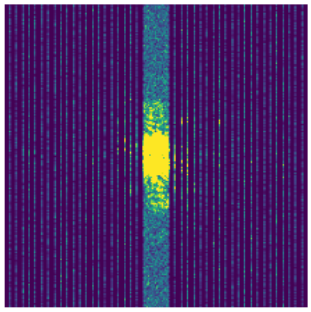
5. Parallel MRI reconstruction

In this section we take real MRI data and perform parallel MRI reconstruction with sensitivity encoding using SPDHG [16, 28]. For our experiments, the data have been sourced from the NYU fastMRI dataset [20, 34].

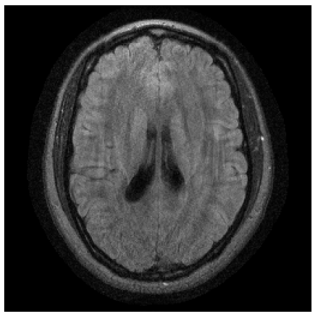
For a system with n coils, we are given n data $b_1, ..., b_n$ which relate to the inverse problems $b_i = A_i x + \eta_i$, where each $A_i : \mathbb{C}^{d_1} \to \mathbb{C}^{d_2}$ is the forward operator from



(a) Coil sensitivities



(b) Data b_1 in k-space



(c) Reconstructed image $\hat{x}$

Figure 1: **Parallel MRI reconstruction with $n = 12$ coils.** (a) Spatial visualization of coil sensitivities for an MRI scanner with 12 electromagnetic coils distributed uniformly in a circle. (b) Data collected by a single coil, encoded in Fourier space. (c) A reconstructed image obtained using sigpy's SenseRecon method [22].

the signal space to the sample space, and $\eta_i \in \mathbb{C}^{d_2}$ represents random noise. The data are undersampled, i.e. $A_i = S \circ F \circ C_i$, where $S : \mathbb{C}^{d_1} \to \mathbb{C}^{d_2}$ is a subsampling operator, $F : \mathbb{C}^{d_1} \to \mathbb{C}^{d_1}$ represents the discrete Fourier transform and $C_i x = c_i \cdot x$ is the element-wise multiplication of x and the i-th coil-sensitivity map $c_i \in \mathbb{C}^{d_1}$. Figure 1a shows the magnitude of the coil sensitivity maps c_i, obtained using sigpy's EspiritCalib method [22] on a dataset collected using $n = 12$ coils, arranged uniformly in a circle around a patient's head.

An image $\hat{x}$ is then reconstructed from the data by solving the regularized least-squares problem

$$\hat{x} \in \arg\min_x \sum_{i=1}^{n} \frac{1}{2}\|A_i x - b_i\|^2 + g(x) \tag{35}$$

where g is a regularizer. We recover our convex minimization template (1) by identifying $\mathbb{C}^d$ with $\mathbb{R}^{2d}$ and setting $X = \mathbb{R}^{2d_1}$, $Y_i = \mathbb{R}^{2d_2}$ and $f_i(y) = \|y - b_i\|^2$.

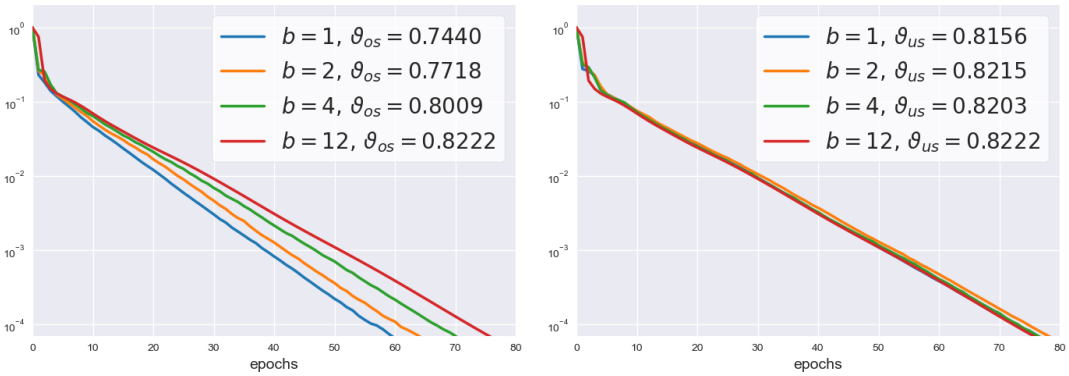
In order to test performance of SPDHG we look into the convergence of its iterations (x^k, y^k) at every epoch. We consider an epoch as the number of iterations required to perform the same amount of computational work (e.g. linear operations) as one iteration of the deterministic PDHG method. In particular, for b-serial and b-nice samplings, where one epoch is roughly equivalent to $m = n/b$ iterations, we define the *relative primal error* and the *relative primal distance* at epoch k, respectively, as

$$\mathbf{e}_b(k) := \frac{\|x^{mk} - \hat{x}\|}{\|\hat{x}\|}, \qquad \mathbf{d}_b(k) := \frac{\|x^{mk} - x^{m(k-1)}\|}{\|x^{m(k-1)}\|},$$

where $\hat{x}$ is a (fixed) solution of (35). In our experiments, $\hat{x}$ is obtained using sigpy's SenseRecon [22] or the deterministic PDHG method [11] for 10^4 iterations.

Similarly, the convergence rate θ from Lemma 6 holds at every iteration k, hence for b-serial and b-nice sampling we can define the *convergence rates per epoch* as

$$\vartheta_{us} = (\theta_{us})^m, \qquad \vartheta_{os} = (\theta_{os})^m, \qquad \vartheta_{un} = (\theta_{un})^m$$



(a) $\mathbf{e}_b$ for optimal b-serial sampling (b) $\mathbf{e}_b$ for uniform b-serial sampling

Figure 2: **Convergence of SPDHG for MRI-L^2.** Relative primal error $\mathbf{e}_b$ for different values of batch size b. The convergence rate is significantly improved when using b-serial sampling with optimized probabilities. A smaller batch size also seems to improve the optimal convergence rate.

with θ_{us}, θ_{os} and θ_{un} defined as in (31) and (34).

All our experiments are implemented in Python using the Operator Discretization Library (ODL) [1]. The code for SPDHG's algorithm is based on the original implementation from [8]. Operator norms such as $\|B\|$ are computed using the power method in ODL.

5.1. L^2 regularization

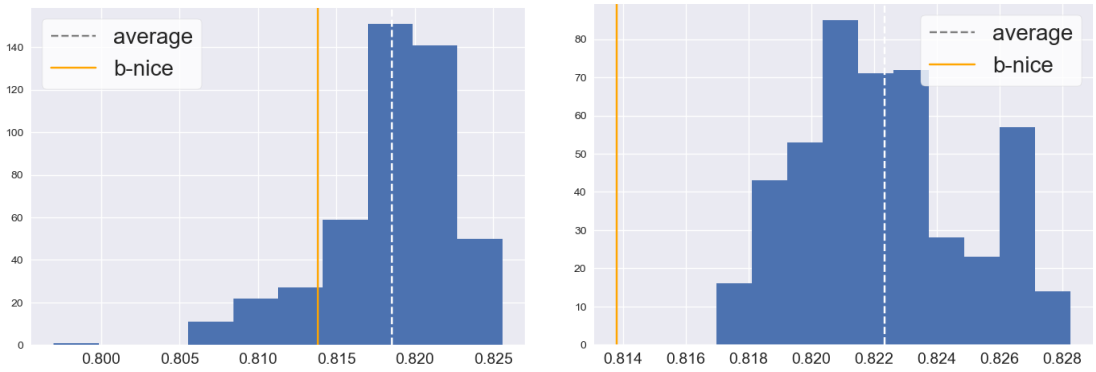
Consider model (35) with L^2 regularizer $g = \frac{\lambda}{2}\|\cdot\|^2$. In this setting, functionals f_i^*, g are strongly convex with convexity parameters $\mu_i = 1$ for all i and $\mu_g = \lambda$, respectively. Hence for both b-serial and b-nice samplings, we can use Lemma 6 to determine the optimal step size parameters τ, σ_i and the optimal rate of convergence θ.

Figure 2 shows the performance of solving (35) with L^2 regularizer and $\lambda = 10^{-2}$ using SPDHG with optimal b-serial and uniform b-serial sampling. For each b it shows the relative primal error $\mathbf{e}_b$ along with its optimal convergence rate per epoch ϑ_{os} or ϑ_{us}. Each curve is the average of 10 independent runs.

The b-serial samplings used in Figures 2a and 2b use the most natural partition of $\{1, ..., n\}$, i.e. the partition consisting of the subsets $I_j = \{(j-1)b, (j-1)b+1, ..., jb\}$ so that

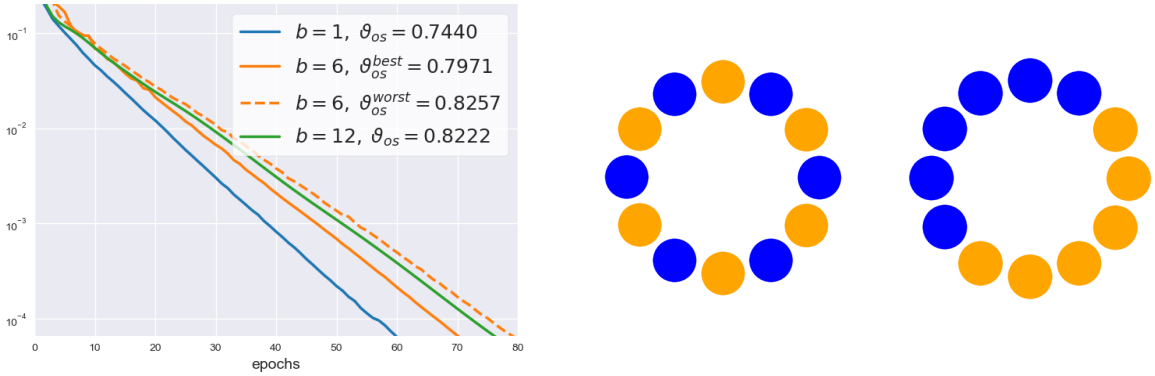
$$\{1, ..., n\} = \bigcup_{j=1}^{m} I_j = \{1, ..., b\} \cup \{b+1, ..., 2b\} \cup ... \cup \{n-b+1, ..., n\}. \tag{36}$$

However, as stated in Section 4, this is not the only way to partition $\{1, ..., n\}$ into m subsets of size b. For instance, from (33) we know the number of partitions for $n = 12$ and $b = 6$ is $\binom{11}{5} = 462$, while for $b = 2$ the number is $\binom{3}{1}\binom{5}{1}\binom{7}{1}\binom{9}{1}\binom{11}{1} = 10\,395$. For each one of these partitions it is possible to compute the optimal convergence rates per epoch ϑ_{os} and ϑ_{us}.



(a) ϑ_{os} (b) ϑ_{us}

Figure 3: **Optimal convergence rates for all partitions of batch size $b = 6$.** Distribution of ϑ_{os} and ϑ_{us} over all $\mathbf{k}(12, 6) = 462$ partitions. The rate ϑ_{un} for uniform b-nice sampling (orange) is better than the average ϑ_{os}, but not better than the rate of the best possible partition. Uniform b-nice performs better than uniform b-serial for any partition.



(a) Best v worst error $\mathbf{e}_b$ (b) Best Partition (c) Worst Partition

Figure 4: **Best and worst partition for optimal serial sampling.** Relative primal error $\mathbf{e}_6$ for the best (solid) and worst (dashed) partitions of batch size $b = 6$. Performance is significantly improved by choosing the right partition. Figures (b) and (c) show the subsets of coils that conform these partitions. Coils that are closer together correspond to more correlated forward operators A_i.

Figure 3 shows the distribution of these rates over all possible partitions for $n = 12, b = 6$. The value of the theoretical convergence rate ϑ_{un} is shown in orange. In Figure 3b we see that uniform b-nice sampling has a better (lower) convergence rate than uniform b-serial sampling with any choice of partition. In Figure 3a, uniform b-nice performs better than the average partition for optimal b-serial, but worse than the best possible partition.

From the values that conform Figure 3a, we have identified the partitions that

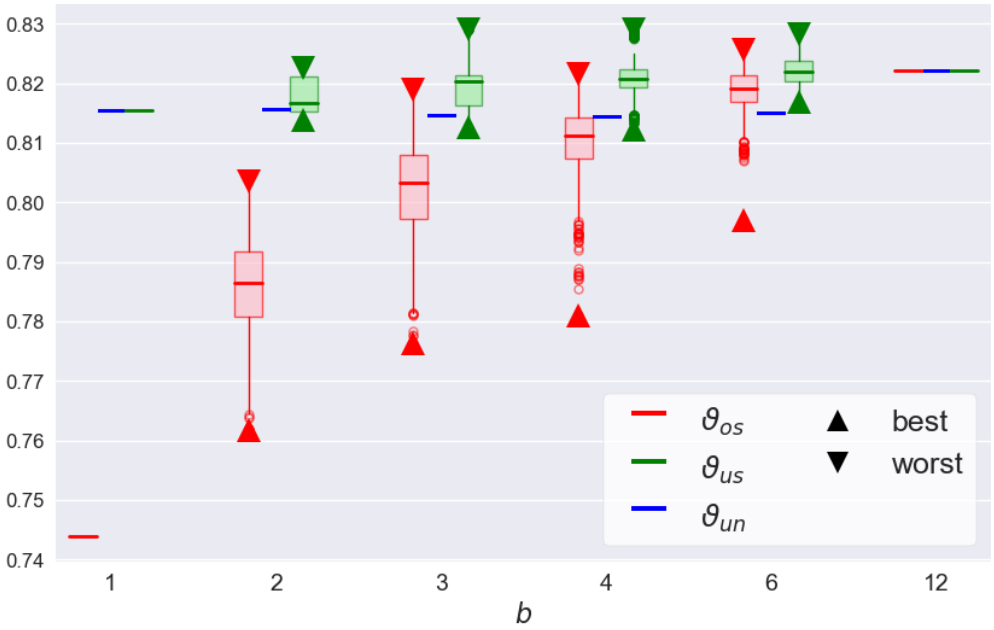


Figure 5: **Distribution of optimal convergence rate ϑ.** Optimal theoretical rates per epoch for optimal b-serial sampling (ϑ_{os}) and uniform b-serial sampling (ϑ_{us}) for all possible partitions of batch size b, for each b. The rate for the best and worst partition is highlighted in each case. Uniform b-nice sampling performs better than uniform b-serial sampling for most partitions, except for the best partition. In all cases, optimal b-serial sampling outperforms the other two samplings when using its best partition.

correspond to the best and worst convergence rate ϑ_{os}. These two partitions are

$$\begin{aligned} \mathbf{P}_{best} &= \{\{1,3,5,7,9,11\},\{2,4,6,8,10,12\}\}, \\ \mathbf{P}_{worst} &= \{\{3,4,5,6,7,8\},\{9,10,11,12,1,2\}\}. \end{aligned} \tag{37}$$

The corresponding coils and their positions are shown in Figures 4b and 4c. The corresponding ϑ_{os} values are $\vartheta_{os}^{best} = 0.7971$ and $\vartheta_{os}^{worst} = 0.8257$. The performance of Algorithm 1 with optimal b-serial sampling using these two partitions is illustrated in Figure 4a. Notice how there is a significant improvement in performance by choosing the correct partition. Furthermore, the error $\mathbf{e}_6$ for optimal b-serial sampling with the worst partition is even worse than the error $\mathbf{e}_n$ for the deterministic PDHG.

Figure 4 also suggests that the physical locations of the coils may determine the best and worst partitions. This seems to confirm the intuitive notion that, since coils that are closer together correspond to more correlated forward operators A_i, the best partition is the one that maximizes correlation between its subsets and minimizes correlation within each subset, i.e. minimizes the value of $\max_j(\|\tilde{A}_j\|)$.

The same experiment can be done to show the distribution of the values ϑ_{os} for all the partitions of size b for other values of b, as well as the distributions of ϑ_{us}. The box plots in Figure 5 show the distributions of ϑ_{os} and ϑ_{us} for different values of b. Trivially, for the special cases of $b = 1$ and $b = n$ there is only one partition possible. The rate

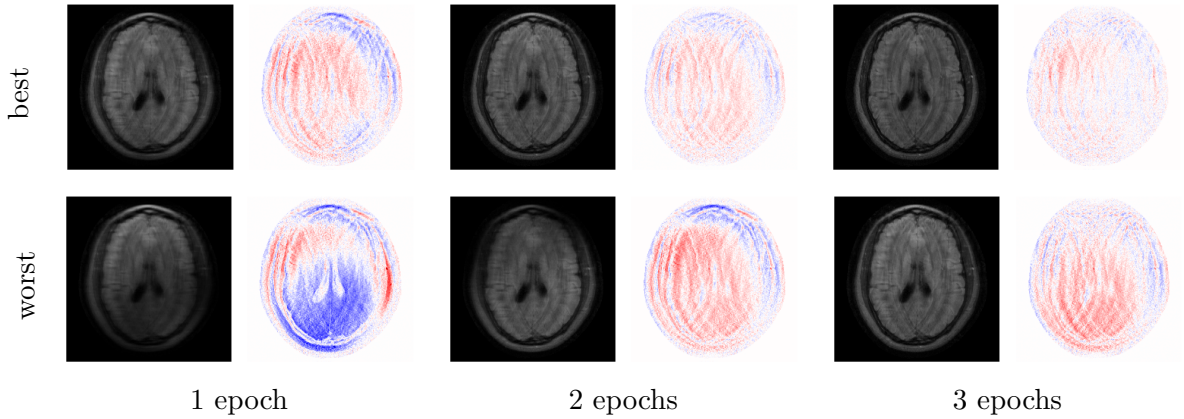


Figure 6: **SPDHG with best and worst sampling.** Reconstructed image $|x^k|$ and real error $\mathrm{Re}(x^k - \hat{x})$, obtained using SPDHG with the best possible sampling, i.e. optimal serial sampling on $b = 1$ (top) and with the worst possible sampling, i.e. uniform b-serial sampling with $b = 3$ on the worst partition (bottom). The performance of SPDHG seems to correspond to the theoretical convergence rates, as suggested in Figure 5.

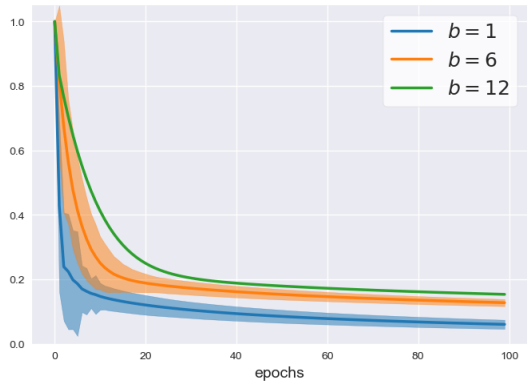
ϑ_{un} for uniform b-nice sampling is independent of any partition.

Notice how for serial sampling ($b = 1$) the values ϑ_{us} and ϑ_{un} coincide, since both samplings have the same probability matrix P, while optimal serial sampling offers significant improvement, simply by using optimal step sizes and probabilities. For $b = n$, all three samplings are identical as they all imply the fully deterministic PDHG, thus $\vartheta_{os} = \vartheta_{us} = \vartheta_{un}$. Notice as well that uniform b-nice performs better than uniform b-serial for most partitions, with the best partition being the exception. For larger values of b, uniform b-nice also seems to perform better than the average partition for optimal b-serial, while for smaller b the average partition for optimal b-serial significantly outperforms uniform b-nice. In all cases, optimal b-serial yields significantly better convergence rates than the other two samplings when using its best possible partition.

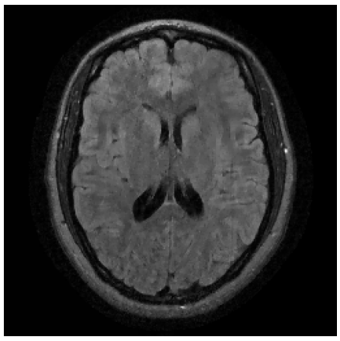
Using the results from Figure 5 we can identify the best and worst possible samplings for SPDHG. The sampling with lowest convergence rate ϑ is attained at $b = 1$ with optimal serial sampling, while the worst rate is given by $b = 3$ with uniform serial sampling with the worst partition. Figure 6 shows the result of using these two different samplings on the same reconstruction problem. One sees that the convergence rates have a clear effect on the efficiency of SPDHG, and that the performance of SPDHG can in general be improved by using a better sampling.

5.2. Total Variation regularization

Consider now the *total variation* (TV) regularizer $g = \lambda\|\nabla \cdot\|_1$ [31]. In our experiments, its proximity operator $\mathrm{prox}_{\tau g}$ is approximated iteratively using the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA) [4].



(a) Mean error $\mathbf{e}_b \pm$ its standard deviation



(b) Solution $\hat{x}$

Figure 7: **SPDHG for MRI-TV with uniform b-serial sampling.** Average relative primal error $\mathbf{e}_b$ plus/minus its standard deviation, out of 10 independent runs. Variation goes to zero as the algorithm converges to a solution. The special case $b = n$ is deterministic and has no variance. The target solution $\hat{x}$ was obtained after 10^4 iterations of PDHG.

For b-serial sampling, we satisfy step-size condition (23) by choosing

$$\tau = \frac{\rho}{\gamma} \qquad \sigma_i = \frac{\rho\gamma\tilde{p}_j}{\|\tilde{A}_j\|^2}, \qquad i \in I_j \; j \in \{1, ..., m\} \tag{38}$$

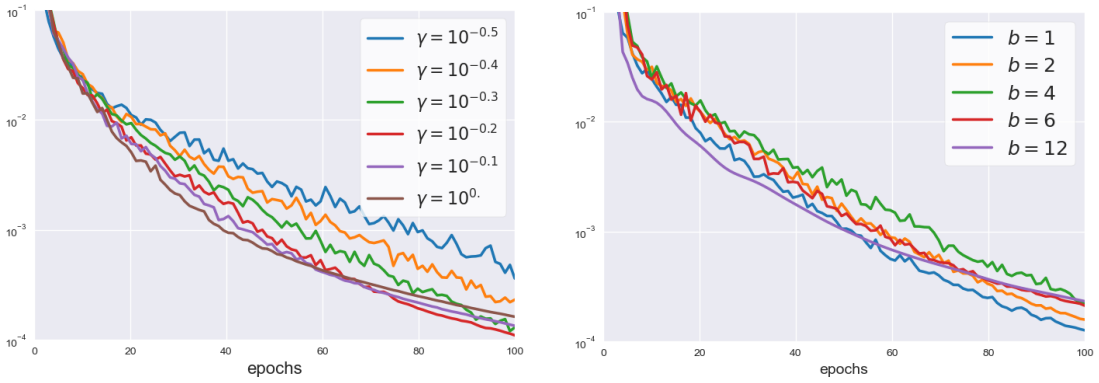
where $\rho \in (0, 1)$ and $\gamma > 0$.

Figure 7a shows the performance of SPDHG applied to (35) with $\lambda = 10^2$, using uniform b-serial sampling and step sizes chosen as in (38) with $\rho = 0.98$ and $\gamma = 1$. The solid lines represent the mean μ of the relative primal error $\mathbf{e}_b$ for 10 different runs of the algorithm, while the shaded color represents the area between $\mu(\mathbf{e}_b) + \sigma(\mathbf{e}_b)$ and $\mu(\mathbf{e}_b) - \sigma(\mathbf{e}_b)$, where $\sigma(\mathbf{e}_b)$ is the standard deviation. The error $\mathbf{e}_b$ in Figure 7a is computed using a solution $\hat{x}$ obtained after 10^4 iterations of the deterministic PDHG algorithm, shown in Figure 7b. Notice how the variance of the iterations decreases over time. In addition, using smaller batch size b shows faster convergence and larger variance at the same time. Trivially, when $b = n$ this is simply the deterministic PDHG method (3) and it has no variance.

In contrast with the L^2 regularization model, the TV regularizer g is not strongly convex and Lemma 6 does not apply. Hence for this model we do not have a formula for determining the optimal step sizes. Instead, for each sampling we test performance for different values of τ, σ_i, chosen as in (38) and (39) with $\rho = 0.98$ and

$$\gamma \in \left\{10^{-2}, 10^{-1.9}, 10^{-1.8}, ..., 10^{1.8}, 10^{1.9}, 10^{2}\right\}.$$

Here, γ represents the trade-off between the primal and the dual step sizes. Figure 8 shows the performance of SPDHG uniform with serial sampling for some values of γ. The choice of step size parameters has an effect in performance even if for all these choices the step size condition (22) is satisfied in exactly the same way, i.e. $\tau\sigma_i\|A_i\|^2 = \rho p_i$ for all values of γ.



(a) Distance $\mathbf{d}_1$ with respect to γ (b) Distance $\mathbf{d}_b$ with respect to b

Figure 8: **SPDHG for MRI-TV with uniform b-serial sampling.** The trade off between step size parameters τ and σ_i, represented here by γ, has an effect on the variance and convergence speed of the algorithm. After establishing an optimal γ_b for each b, we see that smaller batch sizes b appear to yield better performance.

Table 1: **Optimal γ_b value for each b.**

	γ_1	γ_2	γ_3	γ_4	γ_6	γ_{12}
uniform b-serial	$10^{-0.3}$	$10^{-0.5}$	$10^{-0.7}$	$10^{-0.8}$	$10^{-0.8}$	$10^{-1.1}$
uniform b-nice	$10^{-0.2}$	$10^{-0.5}$	$10^{-0.7}$	$10^{-0.8}$	$10^{-0.9}$	$10^{-1.1}$

In this way we can determine a good choice of γ by trying different values and selecting the one that results in the smallest distance $\mathbf{d}_b(k)$ at a fixed time $k = 100$. Doing this for each b we find an "optimal" γ denoted γ_b. The results are listed in Table 1. The same experiment is performed for uniform b-nice sampling, where the step size condition (24) is satisfied by

$$\tau = \frac{\rho}{\gamma} \qquad \sigma = \frac{\rho\gamma b^2}{n^2 \|\mathbb{E}(A_{\mathbb{S}} A_{\mathbb{S}}^*)\|^2} \qquad i \in \{1, ..., n\}. \tag{39}$$

One notices from Table 1 that the optimal γ does depend on b but is more or less constant across the two samplings.

Having established a suitable γ_b for each b, Figure 8b shows the performance of uniform b-serial sampling using γ_b. As in the strongly convex case, here we see an improvement in convergence speed when sampling smaller subsets.

Similarly for uniform b-nice sampling one also finds optimal γ_b. Using these parameters, we can compare the performance of uniform b-serial and uniform b-nice in Figure 9. We see that for most cases uniform b-serial sampling seems to perform just as good or better than uniform b-nice sampling, even when using the naive partition (36).

Indeed, the uniform b-serial sampling used in Figures 7, 8 and 9 use the naive partition from (36). While we do not have theoretical convergence rates to determine what is the best partition, we can try the best and worst partitions (37) from last section

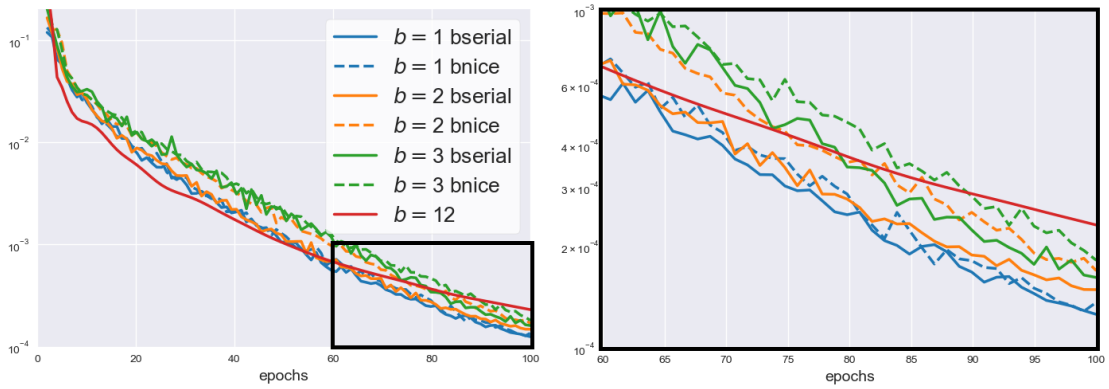


Figure 9: **Uniform b-serial sampling versus uniform b-nice sampling.** Relative primal distance $\mathbf{d}_b$ for SPDHG with uniform b-nice (solid) and uniform b-serial sampling (dashed). Uniform b-serial seems to perform slightly better, even when no information is known about its partition.

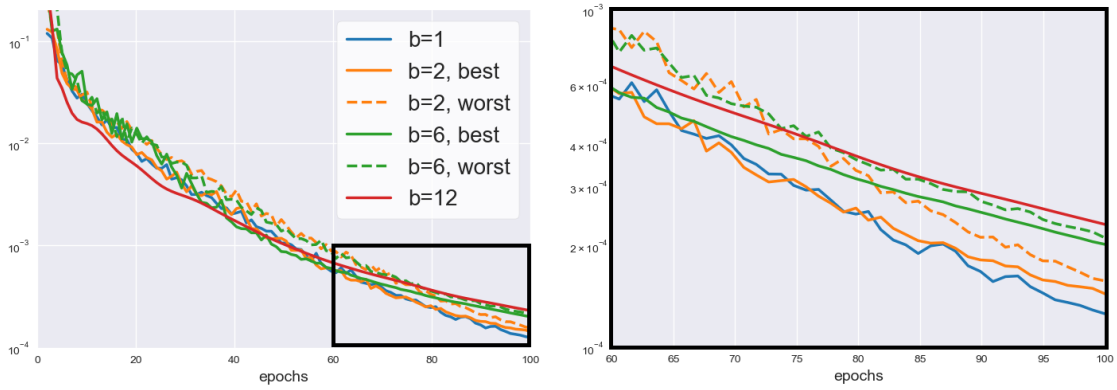


Figure 10: **Uniform b-serial sampling with best and worst partition.** Relative primal distance $\mathbf{d}_b$ for SPDHG using uniform b-serial sampling with partitions $\mathbf{P}_{best}$ and $\mathbf{P}_{worst}$ from (37), represented respectively by a solid and a dashed line. Partition $\mathbf{P}_{best}$ offers better results, even if it was determined for a different model.

as candidates for our TV model.

Figure 10 shows the results of using uniform b-serial sampling with the best and worst partitions as determined for the strongly convex model in the last section. In both cases, we use the same values γ_b for uniform b-serial sampling from Table 1. As before, there is only one possible partition in the cases $b = 1$ and $b = n$. Clearly, the results from using the best partition are better, even when the best and worst partitions were determined using an entirely different model.

This suggests that the right choice of partition, which depends strongly on the numerical properties of the operators $\tilde{A}_j$, may be governed by physical properties of the electromagnetic coils such as their position, as shown in Figure 4.

Similarly, we can test SPDHG with the best and worst samplings from the strongly convex model. Figure 11 shows the performance of SPDHG using these two samplings,

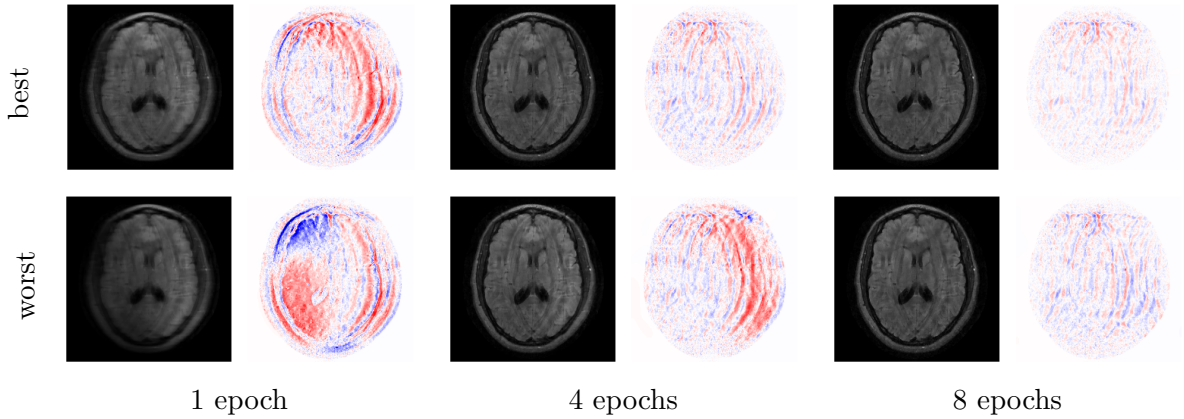


Figure 11: **Uniform b-serial sampling with best and worst partition.** Reconstructed image $|x^k|$ and real error $\mathrm{Re}(x^k - \hat{x})$, obtained using SPDHG with the samplings suggested in Figure 6. Inspired by the strongly convex model, we use uniform serial sampling on $b = 1$ as the best sampling (top) and uniform b-serial sampling with $b = 3$ on the worst partition as the worst possible sampling (bottom). Performance of SPDHG seems improved by the proposed best sampling, even if it was determined for a different model.

as described in Figure 6. The proposed "best" sampling also improves performance in this case, even when determined by the strongly convex model.

6. Conclusions and Discussion

We have closed a gap in the convergence analysis of SPDHG by extending its convergence guarantees to general separable Hilbert spaces and arbitrary sampling. We give a concrete strategy to find parameters that satisfy the required step size condition for all possible random samplings. For several specific random saplings, it is possible to determine theoretically optimal step size parameters. We put these samplings to test using applications to parallel MRI reconstruction.

Our experiments show that the choice of step size parameters directly affects the convergence of the algorithm. Furthermore, the choice of random sampling also strongly influences the performance through several important factors such as whether the indices are chosen independently or grouped into batches (b-nice or b-serial), the partition that defines the batches, the probability with which we sample the batches, and the batch size.

For b-serial sampling, we point out the different ways in which a set can be partitioned into batches. For samplings of fixed batch size, we show that choosing a convenient partition to sample from yields much better performance than the many other possible partitions. Empirically, the spatial location of the coils that define the subsets of this best partition seems to confirm an intuitive notion that coils that are closer together correspond to more correlated forward operators A_i, and hence should

not be grouped together in the same batch, as that yields to worse (smaller) step size parameters.

Finally, partitioning the dual variable into smaller subsets and sampling one at a time can be significantly more efficient than taking larger subsets every iteration. This is the case for the L^2 model with optimal b-serial sampling and for the TV model with any b-serial sampling. For b-nice sampling, in contrast, performance seems to remain consistent across batch size.

In general, our experiments suggest that random samplings perform better than the original deterministic method, b-serial sampling is mostly better than b-nice sampling, and the smaller the sampling, the better. In the strongly convex case serial sampling with with optimal probabilities outperforms all other sampling strategies.

Acknowledgments

MJE and CD acknowledge support from the EPSRC (EP/S026045/1). MJE is also supported by EPSRC (EP/T026693/1), the Faraday Institution (EP/T007745/1) and the Leverhulme Trust (ECF-2019-478). EBG acknowledges the Mexican Council of Science and Technology (CONACyT).

References

[1] Jonas Adler, Holger Kohr, and Ozan Öktem. Operator discretization library (odl), January 2017.

[2] Ahmet Alacaoglu, Olivier Fercoq, and Volkan Cevher. On the convergence of stochastic primal-dual hybrid gradient. *SIAM Journal on Optimization*, 32(2):1288–1318, 2022.

[3] Heinz H Bauschke, Patrick L Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[6] Kristian Bredies and Dirk Lorenz. *Mathematical Image Processing*. Springer, 2018.

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

[8] Antonin Chambolle, Matthias J Ehrhardt, Peter Richtárik, and Carola-Bibiane Schönlieb. Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications. *SIAM Journal on Optimization*, 28(4):2783–2808, 2018.

[9] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[10] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[11] Patrick L Combettes and Jean-Christophe Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[12] Matthias J Ehrhardt, Pawel Markiewicz, and Carola-Bibiane Schönlieb. Faster PET reconstruction with non-smooth priors by randomization and preconditioning. *Physics in Medicine & Biology*, 64(22):225019, 2019.

[13] Ernie Esser, Xiaoqun Zhang, and Tony 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(4):1015–1046, 2010.

[14] Olivier Fercoq, Ahmet Alacaoglu, Ion Necoara, and Volkan Cevher. Almost surely constrained convex optimization. In *International Conference on Machine Learning*, pages 1910–1919. PMLR, 2019.

[15] Olivier Fercoq and Pascal Bianchi. A coordinate-descent primal-dual algorithm with large step size and possibly nonseparable functions. *SIAM Journal on Optimization*, 29(1):100–134, 2019.

[16] Jeffrey A Fessler. Optimization methods for magnetic resonance image reconstruction: Key models and optimization algorithms. *IEEE Signal Processing Magazine*, 37(1):33–40, 2020.

[17] Xiang Gao, Yang-Yang Xu, and Shu-Zhong Zhang. Randomized primal–dual proximal block coordinate updates. *Journal of the Operations Research Society of China*, 7(2):205–250, 2019.

[18] Eric B Gutiérrez, Claire Delplancke, and Matthias J Ehrhardt. Convergence properties of a randomized primal-dual algorithm with applications to parallel mri. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 254–266. Springer, 2021.

[19] William Hager, Cuong Ngo, Maryam Yashtini, and Hong-Chao Zhang. An alternating direction approximate newton algorithm for ill-conditioned inverse problems with application to parallel mri. *Journal of the Operations Research Society of China*, 3(2):139–162, 2015.

[20] Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastmri: A publicly available raw k-space and dicom dataset of knee images for accelerated mr image reconstruction using machine learning. *Radiology: Artificial intelligence*, 2(1), 2020.

[21] Puya Latafat, Nikolaos M Freris, and Panagiotis Patrinos. A new randomized block-coordinate primal-dual proximal algorithm for distributed optimization. *IEEE Transactions on Automatic Control*, 64(10):4050–4065, 2019.

[22] Frank Ong, Jonathan Martin, Will Grissom, Siddharth Srinivasan, Kevin M Johnson, Chris Huynh, Arjun Desai, Zhitao Li, Jon Tamir, Chris Sandino, Efrat Shimron, David Zeng, and Nikolai Mickevicius. mikgroup/sigpy: Minor release to trigger Zenodo for DOI. (v0.1.24). *Zenodo*, 2022
https://doi.org/10.5281/zenodo.5893788.

[23] Frank Ong, Xucheng Zhu, Joseph Y Cheng, Kevin M Johnson, Peder EZ Larson, Shreyas S Vasanawala, and Michael Lustig. Extreme mri: Large-scale volumetric dynamic imaging from continuous non-gated acquisitions. *Magnetic resonance in medicine*, 84(4):1763–1780, 2020.

[24] Andrei Patrascu and Ion Necoara. Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization. *The Journal of Machine Learning Research*, 18(1):7204–7245, 2017.

[25] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. A algorithm for minimizing the Mumford-Shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140, 2009.

[26] Thomas Pock and Antonin Chambolle. Diagonal preconditioning for first order primal-dual algorithms in convex optimization. In *2011 International Conference on Computer Vision*, pages 1762–1769. IEEE, 2011.

[27] Klaas P Pruessmann. Encoding and reconstruction in parallel mri. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 19(3):288–299, 2006.

[28] Klaas P Pruessmann, Markus Weiger, Markus B Scheidegger, and Peter Boesiger. Sense: sensitivity encoding for fast mri. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

[29] Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. *Advances in neural information processing systems*, 28, 2015.

[30] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost

supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

[31] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[32] Georg Schramm and Martin Holler. Fast and memory-efficient reconstruction of sparse poisson data in listmode with non-smooth priors with application to time-of-flight pet. *Physics in Medicine & Biology*, 2022.

[33] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[34] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

[35] Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.