Faster PET Reconstruction with Non-Smooth Priors by Randomization and Preconditioning

Matthias J. Ehrhardt¹, Pawel Markiewicz² & Carola-Bibiane Schönlieb³

¹ Institute for Mathematical Innovation, University of Bath, Bath BA2 7JU, UK
² Centre for Medical Image Computing, London WC1E 6BT, UK
³ Department for Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK

E-mail: m.ehrhardt@bath.ac.uk

December 2018

Abstract. Uncompressed clinical data from modern positron emission tomography (PET) scanners are very large, exceeding 350 million data points (projection bins). The last decades have seen tremendous advancements in mathematical imaging tools, many of which lead to non-smooth (i.e. non-differentiable) optimization problems which are much harder to solve than smooth optimization problems. Most of these tools have not been translated to clinical PET data, as the state-of-the-art algorithms for non-smooth problems do not scale well to large data. In this work, inspired by big data machine learning applications, we use advanced randomized optimization algorithms to solve the PET reconstruction problem for a very large class of non-smooth priors which includes for example total variation, total generalized variation, directional total variation and various different physical constraints. We show on real PET data (FDG and florbetapir) from a Siemens Biograph mMR that about ten projections and backprojections are sufficient for many popular non-smooth priors to produce clinically relevant images; thus showing that the proposed algorithm is fast enough to bring these models into routine clinical practice. Moreover, the proposed algorithm is as fast on the unregularized problem as the clinical standard OSEM but—in contrast to OSEM—is provably convergent, robust and stable for any subset selection.

Keywords: positron emission tomography, convex optimization, randomized optimization, non-smooth optimization, total variation, anatomical priors
1. Introduction

Positron emission tomography (PET) is an important clinical imaging technique as it allows monitoring function of the human body by following a radio-active tracer. The image reconstruction process in PET is challenging as the low number of photon counts call for the Poisson noise modeling and the amount of data is excessively large on modern scanners. While most clinical systems still run non-penalized reconstructions, it has been shown that priors can improve noise control and quantification \[1,2\]. In addition, recent trends suggests that non-smooth priors, such as the total variation \[3\] and its relatives like total generalized variation \[4–6\], are beneficial for imaging applications as they allow smooth variations within regions without oversmoothing sharp boundaries \[3–11\]. Modern PET scanners always come with a second anatomical modality such as computed tomography (CT) or magnetic resonance imaging (MRI). The non-smooth priors can also be used to either incorporate anatomical knowledge from MRI or CT into the reconstruction, e.g. \[11–16\], or to jointly reconstruct PET and the anatomical CT/MRI image \[17–20\]. Only very few algorithms are capable of combining non-smooth priors and the Poisson noise model, e.g. \[7,10,21–23\].

One of the most popular algorithms to solve the resulting non-smooth convex optimization problem is the primal-dual hybrid gradient (PDHG) algorithm\[24,25\]. While this algorithm is flexible enough to solve a variety of non-smooth optimization problems, in every iteration both the projection and the backprojection have to be applied for all projection bins. Moreover, in every iteration computations on vectors that have the size of the data have to be performed. For modern scanners like the Siemens Biograph mMR with span-1 data format, these vectors contain more than 350 million elements and therefore limiting the applicability of this algorithm (and thus many non-smooth priors) to state-of-the-art scanners.

1.1. Contributions

We propose an algorithm, coined Stochastic PDHG or SPDHG for short, which in every iteration performs computations only for a random subset of the data. We show on clinical data from a Siemens Biograph mMR that with this algorithm, for the first time, non-smooth priors become feasible to be used in routine clinical imaging. Numerically, we show that SPDHG is competitive with OSEM on unregularized reconstruction problems but stable with respect to the choice of the subsets due to its mathematically guaranteed convergence. In fact, SPDHG converges to the deterministic solution for any proper subset selection.

In addition to the general randomized solution strategy, we propose two further algorithmic advancements. First, we propose and evaluate the use of data-dependent preconditioners in PDHG and SPDHG for PET image reconstruction. While the convergence theory for a large class of preconditioners has been available since 2011 \[25\],
our proposed preconditioners are the first to be computationally efficient and effective for PET image reconstruction with non-smooth priors. Second, we propose a novel non-uniform sampling strategy, which is necessary to accommodate the differences of data fidelity and regularity. Both randomization and preconditioning can be used independently or can be combined as proposed here in this work.

A few initial findings on randomized reconstruction without preconditioning were published in a conference paper [26].

1.2. PET Reconstruction via Optimization

Given the measured data vector $b \in \mathbb{N}^M$ and the projection model $P$, the PET reconstruction problem can be formulated as the solution to the optimization problem

$$\min_{u \geq 0} \left\{ D(Pu) + \alpha R(u) \right\}$$

(1)

where the data fidelity $D(Pu)$ measures the match of the estimated image $u$ with the data and the prior $\alpha R(u)$ penalizes features that are not desirable in the solution. In other words the prior can be used to avoid solutions which would fit the noisy data too closely. The data fidelity $D$ is (up to constants independent of $u$) the negative log-likelihood of the multi-variate Poisson distribution

$$D(y) = \sum_{i=1}^{M} y_i + r_i - b_i + b_i \log \left( \frac{b_i}{y_i + r_i} \right),$$

with expected value being the sum of the projected image $y$ and the estimated background activity $r$. The latter is needed in order to model non-linear effects such as scatter and randoms. The data fidelity $D$ measures the distance of the estimated data $Pu + r$ to the measured data $b$ in the sense that $D(Pu) \geq 0$ and $D(Pu) = 0$ if and only if $Pu + r = b$. The operator $P$ performs the projection and includes geometric factors, attenuation and normalization.

1.3. Motivating Example: OSEM

If there is no prior, i.e. $\alpha R = 0$, the most common algorithm to solve the optimization problem (1) is the maximum likelihood expectation maximization algorithm (MLEM) [27] defined by

$$u^{k+1} = u^k P^T \left( \frac{b}{Pu^k + r} \right),$$

(2)

where all operations have to be understood element-wise. The computational bottleneck in the MLEM algorithm is the evaluation of the operator $P$ and its transpose $P^T$ in each iteration.
**Figure 1a. Faster with subsets.** Comparison of reconstruction speed of several algorithms. We compare MLEM, OSEM (21 subsets), COSEM (252 subsets) and the proposed SPDHG+ (252 subsets) in terms of PSNR($x^k, x^*$) where $x^*$ is an optimal solution for the florbetapir dataset (see section 4.1) approximated by 5k MLEM iterations. The peak signal-to-noise (PSNR) is defined as $\text{PSNR}(x, x^*) = 20 \log_2(\|x^*\|_\infty / \|x - x^*\|_2)$. The subsets are selected with angles equidistantly divided. OSEM and SPDHG+ are clearly faster than MLEM and COSEM. 1 epoch = 1 full evaluation of the projection operator. *proposed

**Figure 1b. OSEM may become unstable.** OSEM and SPDHG+ are compared for a varying number of subsets. While the speed of SPDHG+ increases with the number of subsets, OSEM fails to converge to the right solution for 100 subsets. *proposed

**Figure 1c. OSEM may become unstable II.** In this example both OSEM and SPDHG+ take 21 subsets with bins equidistantly divided into 21 subsets. In contrast to OSEM, SPDHG+ is robust with respect to this subset selection and achieves a clinically relevant solution.
To overcome this hurdle, it has been proposed to change the update and evaluate the operator and its adjoint only on one out of \( m \) subsets of the data in each iteration. At every iteration \( k \) we choose \( i = \text{mod}(k, m) \) and change update formula (2) to

\[
\tilde{u}^{k+1} = \frac{u^k}{P^T_i} P_i \left( \frac{b_i}{P_i u^k + r_i} \right).
\]

This algorithm became known as ordered subsets expectation maximization (OSEM) [28]. Here \( P_i \) is the restriction of \( P \) onto the \( i \)th subset, i.e. \( P = (P^T_1, \ldots, P^T_m)^T \). While this change of the update equation reduces the computational burden by \( 1/m \), it is in general not guaranteed to converge to a solution of (1), illustrated in Figures 1a and 1b. A convergent version of OSEM, called complete-data OSEM (COSEM), has been developed [29]. While it comes with mathematical convergence guarantees, it is much slower than OSEM (see Figure 1a) and therefore never became popular for the reconstruction of clinical PET data.

MLEM has been extended to include smooth [30] and certain non-smooth [10] prior information, however, conceptually both algorithms intrinsically struggle with the ordered subset acceleration. Also other algorithms have been “accelerated” based on the ordered subset idea, e.g. [31], but are similarly intrinsically unstable due to their non-convergence. See [32] for a numerical comparison and [1] for a validation on clinical PET data. For differentiable priors, a surrogate based technique allows for stable subset acceleration [33–35]. In this work we propose the algorithm SPDHG that is provably convergent and thus stable and robust, see Figures 1b and 1c. SPDHG is flexible enough to be applicable to a large variety of convex and non-smooth priors and is as efficient as OSEM if no explicit prior is being used, see Figures 1a and 1d.

2. Mathematical Model

2.1. Non-Smooth PET Reconstruction with Subsets

As outlined above, PET reconstruction can be formulated in terms of the optimization problem (1). Computationally, it is convenient to rewrite (and solve) the optimization problem (1) in terms of subsets. We denote by \( M \) the number of projection bins. Let
\( \{S_i\} \) be a partition of \([M]\), in the sense that \( \bigcup_{i=1}^{m} S_i = [M] \), where we used the notation \([M] := \{1, \ldots, M\} \). It is not necessary to assume that \( S_i \cap S_j = \emptyset \) for \( i \neq j \). For notational simplicity we will restrict ourselves to the this case. We define

\[
D_i(y) := \sum_{j \in S_i} \varphi_j(y_j)
\]

with the distance function for every data point given by

\[
\varphi(y) := \begin{cases} y + r - b \log(y + r) - b + b \log b & \text{if } y + r \geq 0 \\ \infty & \text{else} \end{cases},
\]

where we omitted the index \( j \) at \( \varphi, y, r \) and \( b \) for readability. Algorithms from convex optimization require the problem to be defined over an entire vector space which we satisfy by extending \( \varphi \) to \( \infty \) for non-positive estimated data \( y + r \). The data and the background are photon counts and therefore have a natural non-negativity constraint.

To allow for the concise notation in (5), we define \( 0 \log 0 := 0 \) and \( -\log 0 := \infty \).

We model the non-negativity constraint for the image \( u \) with the indicator function \( \mathbb{1}_+ \), which is defined as

\[
\mathbb{1}_+(u) = \begin{cases} 0 & \text{if } u \geq 0 \\ \infty & \text{else} \end{cases}.
\]

Thus, this results in the unconstrained optimization problem

**Problem 1 (PET Reconstruction with Subsets).**

\[
u^* \in \arg \min_{u \in \mathbb{R}^N} \left\{ \sum_{i=1}^{m} D_i(P_i u) + \alpha R(u) + \mathbb{1}_+(u) \right\}.
\]

We would like to stress that solving problem (7) is equivalent to solving the original problem (1) for any choice of subsets. In fact, the subset selection becomes a reconstruction parameter that may be varied to speed up the reconstruction procedure.

Often, our prior assumptions involve linear operators, too. One of the most prominent examples of this is the total variation [3]

\[
R(u) = \text{TV}(u) = \|\nabla u\|_{2,1} = \sum_{i} \|\nabla u_i\|_2 = \sum_{i} \left( \sum_{j=1}^{3} (\partial_j u_i)^2 \right)^{1/2},
\]

where we take the 2-norm locally, i.e. at every voxel \( i \) we take the 2-norm of the spatial gradient, and the 1-norm globally, i.e. we sum over all voxels. Forward difference discretization of the gradient operator \( \nabla \) is used as in [21]. Similarly, we use the directional total variation \( R(u) = d\text{TV}(u) = \| \mathbf{D} \nabla u \|_{2,1} \) to incorporate a-priori knowledge about the solution given by an anatomical prior image, see [11, 14, 15, 36] for details.
Solving problem (7) is challenging, even when the involved variables are small and matrix-vector products are easy to compute. The difficulty stems from its non-smoothness. The data term $D_i$ is not finite everywhere and while it is differentiable on its effective domain $\text{dom}(D_i) := \{ y \mid D_i(y) < \infty \}$, the gradient is not globally Lipschitz continuous. In addition, further non-smoothness comes from the constraint $\iota_+$ and the prior $R$ may be non-smooth as well. All of this being said, in PET reconstruction, the variable sizes are actually very large and matrix-vector products expensive to compute.

To apply optimization algorithms to solve (7), we reformulate it as a generic optimization problem of the form

**Problem 2** (Generic Optimization Problem).

$$ x^* \in \arg \min_{x \in X} \left\{ \Psi(x) := \sum_{i=1}^{n} f_i(A_ix) + g(x) \right\}. \quad (8) $$

For instance, for unregularized reconstructions, i.e. $\alpha R = 0$, we may make the association

$$ n = m, \quad g = \iota_+, \quad f_i = D_i, \quad A_i = P_i $$

and reconstructions regularized by the total variation, i.e. $R(u) = \| \nabla u \|_{2,1}$, can be achieved by

$$ n = m + 1, \quad f_i = D_i, i \in [m], \quad f_n = \alpha \| \cdot \|_{2,1} \quad g = \iota_+, \quad A_i = P_i, i \in [m], \quad A_n = \nabla. \quad (9) $$

### 2.2. Optimization with Saddle-Point Problems

Instead of solving problem (8) directly, it is more efficient to reformulate the minimization problem as a saddle point problem making use of the **convex conjugate** of a functional, see e.g. [37].

**Definition 1** (Convex Conjugate). Let $f : Y \rightarrow \mathbb{R}_\infty := \mathbb{R} \cup \{ \infty \}$ be a functional with extended real values. Then we define the convex conjugate of $f$ as $f^* : Y \rightarrow \mathbb{R}_\infty$ with

$$ f^*(y) = \sup_{x} \{ \langle y, x \rangle - f(x) \}. $$

For convex, proper and lower semi-continuous (lsc) functionals $f$ we have that $f^{**} = f$, see e.g. [37], and thus $f(x) = \sup_y \{ \langle x, y \rangle - f^*(y) \}$. Then, with $Y = \prod_{i=1}^{n} Y_i$, problem (8) is equivalent to

**Problem 3** (Generic Saddle Point Problem).

$$ \min \sup_{x \in X} \left\{ \sum_{i=1}^{n} \langle A_ix, y_i \rangle - f^*_i(y_i) + g(x) \right\}. \quad (10) $$

We will refer to the variable $x$ as the **primal variable** and to $y$ as the **dual variable**.
Example 1. The convex conjugate of the PET distance function \( D^*_i(y) = \sum_{j \in S_i} \varphi^*_j(y_j) \) with

\[
\varphi^*(y) = \begin{cases} 
-yr - b \log(1 - y) & \text{if } y \leq 1 \\
\infty & \text{else}
\end{cases}
\]  

(11)

(9)

(10)

where we omitted the index \( j \) at \( \varphi, y, b \) and \( r \) for readability.

The derivation of the formulas in this and the following example are omitted for brevity.

As some (or all) of the \( f_i \) and \( g \) in (8) are non-smooth, we make use of the proximity operator of these. Our definition varies slightly from the usual definition as we allow the step size parameter to be matrix-valued. For a symmetric and positive definite matrix \( S \), we define the weighted norm \( \|x\|_S \) as

\[
\|x\|_S^2 = \langle S^{-1/2}x, S^{-1/2}x \rangle.
\]

Definition 2 (Proximal Operator). Let \( S \) be a symmetric and positive definite matrix. Then we define the proximal operator of \( f \) with metric (or step size) \( S \) as

\[
\text{prox}_S f(x) := \arg \min_z \{ \|z - x\|_S^2 + f(z) \}.
\]

From here on, \( S \) and \( T \) will always be diagonal, symmetric and positive definite matrices.

Example 2. The proximity operator of the non-negativity constraint (6) is given element-wise by

\[
\text{prox}_T \tilde{i}(x) = \max(x, 0).
\]

Example 3. Let \( S_i = \text{diag}(\sigma_j)_{j \in S_i} \). The proximal operator of the convex conjugate of the PET distance (11) can be computed element-wise as \( \text{prox}_{D^*_i}(y) = \text{prox}_{\varphi^*_j}(y_j) \). For each element, the proximal operator is given by

\[
\text{prox}_{\varphi^*_j}(y) = \frac{1}{2} \left[ w + 1 - \left( (w - 1)^2 + 4\sigma b \right)^{1/2} \right],
\]

where we again omitted the indices \( j \) for readability and denoted \( w = y + \sigma r \).

3. Algorithm

The saddle point problem (10) (and therefore the PET reconstruction problem (8)) can be solved with the PDHG [21], see Algorithm 1. It consists of very simple operations involving only basic linear algebra, matrix-vector multiplications and the evaluations of proximal operators. As seen in line 4 of the pseudo-code, PDHG updates all dual variables simultaneously. Therefore, in line 4 and 5, the projection and backprojection that corresponds to the whole data set have to be evaluated. The idea of SPDHG, Algorithm 2, is to only select one dual variable randomly in each iteration (line 4) and to perform the update accordingly (line 5 and 6). An important detail is the extrapolation in line 8 with the inverse of the probability \( p_i \) that \( i \) will be selected in each iteration. This guarantees the convergence as proven in Theorem 1 below.
Algorithm 1 Primal-Dual Hybrid Gradient (PDHG) to solve (10). Default values given in brackets.

**Input:** iterates \(x(= 0), y(= 0)\), step parameters \(S = \{S_i\}, T\)

1: \(z = z = A^T y (= 0)\)
2: for \(k = 1, \ldots \) do
3: \(x = \text{prox}_{g}(x - Tz)\)
4: \(y_i^+ = \text{prox}_{S_i}(y_i + S_i A_i x)\) for \(i = 1, \ldots, n\)
5: \(\Delta z = \sum_{i=1}^{n} A_i^T (y_i^+ - y_i)\)
6: \(z = z + \Delta z, y = y^+\)
7: \(z = z + \Delta z\)

Algorithm 2 Stochastic Primal-Dual Hybrid Gradient (SPDHG) to solve (10). Default values given in brackets.

**Input:** iterates \(x(= 0), y(= 0)\), step parameters \(S = \{S_i\}, T\)

1: \(z = z = A^T y (= 0)\)
2: for \(k = 1, \ldots \) do
3: \(x = \text{prox}_{g}(x - Tz)\)
4: Select \(i \in [n]\) at random with probability \(p_i\)
5: \(y_i^+ = \text{prox}_{S_i}(y_i + S_i A_i x)\)
6: \(\Delta z = A_i^T (y_i^+ - y_i)\)
7: \(z = z + \Delta z, y_i = y_i^+\)
8: \(z = z + \frac{1}{p_i} \Delta z\)

3.1. Convergence

SPDHG is guaranteed to converge for any \(f_i\) and \(g\) which are convex, proper and lsc.

We now state a very general convergence result which can be derived from [38, Theorem 4.3]. The actual proof is omitted here for brevity. For more details on convergence and convergence rates we refer the reader to [38].

**Theorem 1** (Convergence). Assume that the sampling is proper, i.e. the probability \(p_i\) for an index \(i \in [n]\) to be sampled is positive. Let the step length parameters \(T = \min_{i \in [n]} T_i, S_i\) be chosen such that for all \(i \in [n]\) the following bound on the operator norm holds:

\[
\left\| S_i^{1/2} A_i T_i^{1/2} \right\|^2 < p_i \tag{12}
\]

Then for any initialization, the iterates \((x, y)\) of SPDHG (Algorithm 2) converge (in a generalized sense) to a saddle point of (10) almost surely in a Bregman distance.

**Remark 1** (Computational Efficiency). Each iteration of Algorithm 2 is computationally efficient as only projections and backprojections corresponding to the randomly selected subset \(i\) of the data are required. However, the algorithm maintains the whole
backprojected dual variable \( z = P^T y = \sum_{i=1}^{m} P_i^T y_i \) and in each iteration updates the primal variable with it.

**Remark 2** (Memory Requirements). The memory requirement of Algorithm 2 is higher compared to OSEM or gradient descent but still reasonably low. It requires memory equivalent to two images \((z, \overline{z})\) and up to twice the binned sinogram data \((y, y^+)\) in addition to the necessary memory consumption (output image, sinogram data, background).

**Remark 3** (Sampling). SPDHG allows any kind of random selection as long as the draws are independent and the probability that block \( i \) is being selected with positive probability \( p_i > 0 \). We will investigate two choices of sampling in the numerical section of this paper. A more thorough numerical and theoretical investigation will be subject of future work.

### 3.2. Step Sizes and Preconditioning

We will now discuss two different choices of step sizes under which SPDHG is guaranteed to converge. The proof of the following theorem uses arguments from [38] and [25] and is omitted here for brevity.

**Theorem 2** (Step Size Parameters). Let \( \rho < 1 \). Then, condition (12) of Theorem 1 is satisfied by

\[
S_i = \frac{\rho}{\|A_i\|} I, \quad T_i = \frac{\rho p_i}{\|A_i\|} I. \tag{13}
\]

Moreover, if \( A_i \) has only non-negative elements, then condition (12) is also satisfied by

\[
S_i = \text{diag}\left(\frac{\rho}{A_i^{1/2}}\right), \quad T_i = \text{diag}\left(\frac{\rho p_i}{A_i^{1/2}}\right). \tag{14}
\]

An example of preconditioned step sizes (14) is shown in Figure 2.

**Remark 4.** If \( n = 1 \) and \( p_i = 1 \), then the step sizes (13) can be identified with the scalar step sizes \( \sigma_i = \rho/\|A_i\| \) and \( \tau = \rho/\|A_i\| \) which are commonly chosen for PDHG.

**Remark 5.** Note that the non-negativity condition holds for the PET projection operator (and any other ray tracing based operator). Moreover, the step size \( T \) in (14) resembles the sensitivities used in the update of MLEM (2) and OSEM (3). In addition, a similar preconditioning is performed for the dual variable in the data space.

### 4. Numerical Results

The numerical experiments use the open-source package ODL [39] which allows for efficient algorithm prototyping in Python. The projection and backprojections are computed with CUDA through the open-source package NiftyPET [40] and is accessible via Python. All results in this section were obtained by selecting subsets with
Faster PET Reconstruction by Randomization and Preconditioning

4.1. Data

We validate the numerical performance of the proposed algorithm on two clinical PET data sets which we refer to as FDG and florbetapir. The two separate PET brain datasets each use a distinct radiotracer: $[^{18}\text{F}]$FDG for epilepsy and $[^{18}\text{F}]$florbetapir for the neuroscience sub-study Insight’46 of the Medical Research Council National Survey of Health and Development [41]. The epileptic patient was injected with 250 MBq of FDG, one hour before the 15-minute PET acquisition. The neuroscience volunteer was injected with 370 MBq of florbetapir and scanned dynamically for one hour, starting at the injection time. The last ten minutes were used as a measurement of amyloid deposition, which for the participant was negative.

4.2. Results for Total Variation

In this section we analyze the impact of various choices within SPDHG on its performance, from randomness over sampling to preconditioning. The test case is total variation prior as in (9).

4.2.1. Randomness

Figure 3 shows the effect of randomness where we compare the deterministic PDHG to SPDHG with uniform sampling and scalar step sizes (13) for two different number of subsets. The horizontal axis reflects the number of projections in equidistantly divided angles. Moreover, all PSNR or relative objective comparisons are performed by first computing an approximate minimizer $x^*$ by the deterministic PDHG. The Python code and one data set will be made accessible upon acceptance of this manuscript.
each algorithm, we call one full projection for the whole data one “epoch”. Here and in
the following dashed lines represent deterministic and solid lines randomized algorithms.
We can easily see that both random variants are faster than their deterministic PDHG.
Moreover, the randomized SPDHG becomes faster by choosing a larger number of
subsets.

![Figure 3. Deterministic v randomized.](image)
The results for the data set FDG with TV prior show that the randomized algorithms are much faster than their deterministic counterpart. Moreover, more subsets leads to a faster algorithm.

4.2.2. Sampling
The effect of different choices of sampling is shown in Figure 4. We compare two different samplings: uniform sampling and balanced sampling. The uniform sampling chooses all indices \(i \in [n]\) with equal probability \(p_i = 1/n\). In contrast, for balanced sampling we choose with uniform probability either data or prior. If we choose data, then we select a subset again randomly with uniform probability. Thus, the probability for each subset of the data to be selected is \(p_i = 1/(2m), i \in [m]\) and for the prior to be selected \(p_n = 1/2\).

We make two observations. First, balanced sampling is always faster than uniform sampling. This shows the importance of updating the dual variable associated to the prior. Second, for either sampling choosing a larger number of subsets again improves the performance.

4.2.3. Preconditioning
As shown in Theorem 2, the step size parameters \(T\) and \(S_i\) can be chosen either as scalars (13) or as vectors (14), the latter can be seen as a form of preconditioning. Results are shown in Figure 5 where we see that preconditioning may accelerate the convergence of either the deterministic PDHG or the randomized SPDHG. Moreover, combining randomization and preconditioning yields an even faster algorithm. The relative objective value is defined as \((\Psi(x^k) - \Psi(x^*)) / (\Psi(x^0) - \Psi(x^*))\).

4.2.4. Performance of Proposed Algorithm
Based on the previous three examples, we propose to combine randomization, balanced sampling and preconditioning, which we
Figure 4. Uniform v balanced sampling. In addition to increasing the number of subsets, the sampling is also very important for the speed of the algorithm: 21 subsets with balanced sampling is faster than 100 subsets with uniform sampling.

Figure 5. Preconditioning can be used with and without randomization. The preconditioned algorithms are much faster than without preconditioning.

refer to as SPDHG+. Figure 6 shows the visual performance of PDHG and SPDHG+. In contrast to the deterministic PDHG, the proposed SPDHG+ yields a clinically relevant image after only 10 epochs.

Figure 6. Qualitative results show that in contrast to the deterministic PDHG, the proposed SPDHG+ (252 subsets) yields a clinically relevant image after only 10 epochs.
4.3. Further Numerical Results

4.3.1. Anisotropic Total Variation Anisotropic total variation decouples the penalization of the derivatives. The mathematical model is similar to the isotropic TV model (9), the only difference being the norm how the total variation is measured: \( f_n = \alpha \| \cdot \|_{1,1} \). It can be seen in Figure 7 for \texttt{florbetapir} that with randomization and preconditioning only a few epochs are needed to reconstruct clinically relevant images.

![Fig 7. Anisotropic TV regularized reconstruction from FDG data. Top: PDHG and SPDHG+ (252 subsets) reconstructions after 10 epochs. Bottom: Quantitative results show a significant speed-up from randomization and preconditioning. Increasing the number of subsets from 21 to 252 has little effect on this data set.](image)

4.3.2. Directional Total Variation Anatomical information from a co-registered MRI is available on combined PET-MR scanners. The structural information of the anatomy can be utilized by the directional total variation prior, see [11, 14, 15, 36] for details. The mathematical model is similar to the total variation model (9), except for an additional matrix \( \mathbf{D} \). Thus, the only difference is \( \mathbf{A}_n = \mathbf{D} \nabla \). A numerical example is shown in Figure 8 for the data set \texttt{florbetapir}.

![Fig 8.](image)

4.3.3. Total Generalized Variation More sophisticated regularization can be achieved by the total generalized variation (TGV) \([5, 6]\)

\[
\text{TGV}_{\alpha_0, \alpha_1}(u) = \inf_w \{ \alpha_0 \| \nabla u - w \|_{2,1} + \alpha_1 \| \mathcal{E} w \|_{2,1} \}
\]

which can balance first and second order regularization and achieves edge-preserved reconstruction while avoiding the stair-casing artifact. We can solve the TGV...
regularized PET reconstruction problem by solving problem (8) with the assignment $x = (u, w)$ and

$$n = m + 2, \quad A_i = (P_i, 0), i \in [m], \quad A_{m-1} = (\nabla, -I), \quad A_n = (0, \mathcal{E})$$

$$g(x) = \iota(u), \quad f_i = D_i, i \in [m], \quad f_{n-1} = \alpha_0 \| \cdot \|_{2,1}, \quad f_n = \alpha_1 \| \cdot \|_{2,1},$$

where $\mathcal{E}$ is a symmetrized gradient operator, see [5, 6] for more details.

The numerical results shown in Figure 9 are in line with the previous findings indicating that randomization and preconditioning can significantly speed up the reconstruction. However, we notice a significant increase in performance by increasing the number of subsets from 21 to 252.

4.3.4. Comparison of Mathematical Models

We conclude this section by a comparison of various methods on both data sets in Figures 10 and 11. While we leave the detailed visual comparisons to the reader, we would like to note that all these images use the same number of projections so have basically the same computational cost.

5. Discussion

The extensive numerical experiments all consistently confirm that randomization and preconditioning both speed up the reconstruction. These trends were irrespective of the data set and the chosen prior. The convergence speed in our work was abstractly defined by a solution of the underlying mathematical optimization model approximated...
with way too many iterations than would be feasible in routine clinical practice. This strategy was chosen intentionally as we did not want to target a specific clinical use case. After these successful initial trials, in the future we will collaborate with medical researchers and clinicians to focus on specific use cases where each use case defines its own metric of what images we wish to reconstruct.

The focus of this contribution was on non-smooth priors like total variation and its descendants like total generalized variation and directional total variation. However, as long as the proximal operators are simple to evaluate, the proposed randomized and preconditioned algorithm can be applied to any other model, too. It would be of interest to compare this algorithm to convergent subset accelerated algorithms for smooth priors like BSREM \cite{33,34}, TRIOT \cite{35} and OS-SPS \cite{34}.

We highlighted the improvements from choosing different distributions for subset selection by comparing “uniform” and “balanced sampling”. Further improvements are expected by optimizing the probability selection of this algorithm. This can either be an optimal distribution that is constant along the iterations or even developing over the course of the iterations. We will investigate this direction further in the future.
Figure 10. Comparison of several reconstruction approaches for the FDG data. All approaches have about the same computational cost.

Figure 11. Comparison of several reconstruction approaches for the florbetapir data. All approaches have about the same computational cost.
With the exception of Figures 7 and 8 where 21 and 252 subsets were similarly fast, more subsets always resulted in a faster algorithm. There are neither theoretical nor numerical insights how the speed will depend on the subset selection and if more subsets always result in a faster algorithm. However, the numerical evidence suggests that increasing the number of subsets never decreases the speed of the algorithm. This being said, due to the per iteration computational costs, from a practical point of view, there will be an optimal number of subsets that might depend on the prior and even the data (e.g. number of counts) to be reconstructed.

Moreover, the algorithm does not exploit any special structure of our optimization problem like smoothness or strong convexity. It is likely that exploiting these properties will lead to additional speed-up. However, as these properties for the PET data term depend on the acquired data, it is unlikely that a straightforward approach will be sufficient and a tailored solution will be necessary.

6. Conclusion

We introduced a convergent subset accelerated algorithm for the reconstruction of PET images with non-smooth priors. The algorithm was enhanced by data-dependent preconditioning. Our numerical results showed that using both randomized subset selection and preconditioning can dramatically speed up the convergence of an iterative reconstruction algorithm. It was observed that a computational effort similar to the current clinical standard OSEM was sufficient for many non-smooth priors, showing that these are now, for the first time, feasible to be used in daily clinical routine.

While this observations were consistent among two data sets with different tracers, more studies are needed to confirm the benefits of this reconstruction strategy. Overall, this algorithmic advancement has the potential to hugely impact the PET reconstruction landscape as advanced mathematical models can now be combined with efficient and convergent subset acceleration.

Acknowledgments

M.J.E. and C.-B.S. acknowledge support from Leverhulme Trust project “Breaking the non-convexity barrier”, EPSRC grant “EP/M00483X/1”, EPSRC centre “EP/N014588/1”, the Cantab Capital Institute for the Mathematics of Information, and from CHiPS and NoMADS (Horizon 2020 RISE project grants). In addition, M.J.E. acknowledges support from the EPSRC platform grant “EP/M020533/1”. Moreover, C.-B.S. is thankful for support by the Alan Turing Institute. In addition, all authors gratefully acknowledge the hardware donation by the NVIDIA Corporation.

References

[1] E. J. Teoh et al., “Phantom and Clinical Evaluation of the Bayesian Penalized
REFERENCES

Likelihood Reconstruction Algorithm Q.Clear on an LYSO PET/CT System,” *J Nucl Med*, vol. 56, no. 9, pp. 1447–1452, 2015.

[2] S. Ahn et al., “Quantitative Comparison of OSEM and Penalized Likelihood Image Reconstruction using Relative Difference Penalties for Clinical PET,” *Phys Med Biol*, vol. 60, no. 15, pp. 5733–5751, 2015.

[3] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear Total Variation based Noise Removal Algorithms,” *Physica D*, vol. 60, no. 1, pp. 259–268, 1992.

[4] M. Benning, P. Heins, and M. Burger, “A Solver for Dynamic PET Reconstructions based on Forward-Backward-Splitting,” in *AIP Conference Proceedings*, vol. 1281, pp. 1967–1970, 2010.

[5] K. Bredies, K. Kunisch, and T. Pock, “Total Generalized Variation,” *SIAM J Imaging Sci*, vol. 3, no. 3, pp. 492–526, 2010.

[6] K. Bredies and M. Holler, “A TGV-Based Framework for Variational Image Decompression, Zooming, and Reconstruction. Part II: Numerics,” *SIAM J Imaging Sci*, vol. 8, no. 4, pp. 2851–2886, 2015.

[7] S. Setzer, G. Steidl, and T. Teuber, “Deblurring Poissonian Images by Split Bregman Techniques,” *Journal of Visual Communication and Image Representation*, vol. 21, no. 3, pp. 193–199, 2010.

[8] S. Anthoine, J.-F. Aujol, Y. Boursier, and C. Melot, “Some Proximal Methods for Poisson Intensity CBCT and PET,” *Inverse Problems and Imaging*, vol. 6, no. 4, pp. 565–598, 2012.

[9] M. Burger and S. Osher, “A Guide to the TV Zoo,” in *Level Set and PDE Based Reconstruction Methods in Imaging*, vol. 2090 of *Lecture Notes in Mathematics*, pp. 1–70, Springer, 2013.

[10] A. Sawatzky et al., “EM-TV Methods for Inverse Problems with Poisson Noise,” in *Level Set and PDE Based Reconstruction Methods in Imaging*, vol. 2090 of *Lecture Notes in Mathematics*, Springer, 2013.

[11] M. J. Ehrhardt and M. M. Betchke, “Multi-Contrast MRI Reconstruction with Structure-Guided Total Variation,” *SIAM J Imaging Sci*, vol. 9, no. 3, pp. 1084–1106, 2016.

[12] J. E. Bowsher et al., “Utilizing MRI Information to Estimate F18-FDG Distributions in Rat Flank Tumors,” in *IEEE NSS-MIC*, pp. 2488–2492, 2004.

[13] M. Hintermüller, M. Holler, and K. Papafitsoros, “A Function Space Framework for Structural Total Variation Regularization with Applications in Inverse Problems,” *Inverse Probl*, vol. 34, no. 6, p. 064002, 2018.

[14] M. J. Ehrhardt et al., “PET Reconstruction with an Anatomical MRI Prior using Parallel Level Sets,” *IEEE Trans Med Imaging*, vol. 35, no. 9, pp. 2189–2199, 2016.

[15] G. Schramm et al., “Evaluation of Parallel Level Sets and Bowsher’s Method as Segmentation-Free Anatomical Priors for Time-of-Flight PET Reconstruction,” *IEEE Trans Med Imaging*, vol. 0062, no. c, pp. 1–1, 2017.
REFERENCES

[16] A. Mehranian et al., “PET Image Reconstruction using Multi-Parametric Anato-
Functional,” Phys Med Biol, 2017.

[17] M. J. Ehrhardt et al., “Joint Reconstruction of PET-MRI by exploiting Structural
Similarity,” Inverse Probl, vol. 31, p. 015001, 2015.

[18] F. Knoll et al., “Joint MR-PET Reconstruction using a Multi-Channel Image
Regularizer,” IEEE Trans Med Imaging, vol. 36, no. 1, 2016.

[19] J. Rasch, E.-M. Brinkmann, and M. Burger, “Joint Reconstruction via Coupled
Bregman Iterations with Applications to PET-MR Imaging,” Inverse Probl, vol. 34,
no. 1, p. 014001, 2018.

[20] A. Mehranian et al., “Synergistic PET and SENSE MR Image Reconstruction using
Joint Sparsity Regularization,” IEEE Trans Med Imaging, vol. 37, no. 1, pp. 20 –
34, 2018.

[21] A. Chambolle and T. Pock, “A First-Order Primal-Dual Algorithm for Convex
Problems with Applications to Imaging,” J Math Imaging Vision, vol. 40, no. 1,
pp. 120–145, 2011.

[22] F.-X. Dupe, M. J. Fadili, and J.-L. Starck, “Inverse Problems with Poisson Noise:
Primal and Primal-Dual Splitting,” in IEEE ICIP, pp. 1901–1904, 2011.

[23] M. A. T. Figueiredo and J. M. Bioucas-Dias, “Frame-Based Deconvolution of
Poissonian Images using Alternating Direction Optimization,” in IEEE ICIP,
vol. 19, pp. 3549–3552, 2010.

[24] 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
J Imaging Sci, vol. 3, no. 4, pp. 1015–1046, 2010.

[25] T. Pock and A. Chambolle, “Diagonal Preconditioning for First Order Primal-Dual
Algorithms in Convex Optimization,” in IEEE ICCV, pp. 1762–1769, 2011.

[26] M. J. Ehrhardt et al., “Faster PET Reconstruction with a Stochastic Primal-Dual
Hybrid Gradient Method,” in SPIE Optics+Photonics:Wavelets and Sparsity XVII,
(San Diego), 2017.

[27] L. A. Shepp and Y. Vardi, “Maximum Likelihood Reconstruction for Emission
Tomography,” IEEE Trans Med Imaging, vol. 1, no. 2, pp. 113–22, 1982.

[28] H. M. Hudson and R. S. Larkin, “Accelerated Image Reconstruction Using Ordered
Subsets of Projection Data,” IEEE Trans Med Imaging, vol. 13, no. 4, pp. 601–609,
1994.

[29] I.-t. Hsiao, A. Rangarajan, and G. Gindi, “A Provably Convergent OS-EM Like
Reconstruction Algorithm for Emission Tomography,” in Proceedings of SPIE,
vol. 4684, pp. 10–19, 2002.

[30] P. J. Green, “Bayesian Reconstructions from Emission Tomography Data using a
Modified EM Algorithm,” IEEE Trans Med Imaging, vol. 9, no. 893, 1990.
[31] M. G. McGaffin and J. A. Fessler, “Alternating Dual Updates Algorithm for X-ray CT Reconstruction on the GPU,” *IEEE Trans Comput Imaging*, vol. 1, no. 3, pp. 186–199, 2015.

[32] L. Cheng, E. Asma, S. Ahn, and R. M. Manjeshwar, “Comparison of Numerical Convergence Speeds of Convergent and Accelerated Algorithms for Penalized Likelihood PET Image,” in *IEEE NSS-MIC*, pp. 3–6, 2013.

[33] A. R. De Pierro and M. E. B. Yamagishi, “Fast EM-like Methods for Maximum ”A Posteriori” Estimates in Emission Tomography,” *IEEE Trans Med Imaging*, vol. 20, no. 4, pp. 280–288, 2001.

[34] S. Ahn and J. A. Fessler, “Globally Convergent Image Reconstruction for Emission Tomography using Relaxed Ordered Subsets Algorithms,” *IEEE Trans Med Imaging*, vol. 22, no. 5, pp. 613–626, 2003.

[35] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero, “Convergent Incremental Optimization Transfer Algorithms: Application to Tomography,” *IEEE Trans Med Imaging*, vol. 25, no. 3, pp. 283–296, 2006.

[36] L. Bungert et al., “Blind Image Fusion for Hyperspectral Imaging with the Directional Total Variation,” *Inverse Probl*, vol. 34, no. 4, p. 044003, 2018.

[37] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2011.

[38] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schönlieb, “Stochastic Primal-Dual Hybrid Gradient Algorithm with Arbitrary Sampling and Imaging Applications,” *SIAM J Optim*, vol. 28, no. 4, pp. 2783–2808, 2018.

[39] J. Adler, H. Kohr, and O. Öktem, “Operator Discretization Library (ODL),” Jan. 2017.

[40] P. J. Markiewicz et al., “NiftyPET: a High-throughput Software Platform for High Quantitative Accuracy and Precision PET Imaging and Analysis,” *Neuroinformatics*, vol. 16, no. 1, pp. 95–115, 2018.

[41] C. A. Lane et al., “Study Protocol: Insight 46 - A Neuroscience Sub-Study of the MRC National Survey of Health and Development,” *BMC Neurology*, vol. 17, no. 1, pp. 1–25, 2017.