Faster PET reconstruction with a stochastic primal-dual hybrid gradient method

Matthias J. Ehrhardt, Pawel Markiewicz, Antonin Chambolle, Peter Richtárik, Jonathan Schott, et al.


Event: SPIE Optical Engineering + Applications, 2017, San Diego, California, United States
Faster PET Reconstruction with a Stochastic Primal-Dual Hybrid Gradient Method

Matthias J. Ehrhardt\textsuperscript{a}, Pawel Markiewicz\textsuperscript{b}, Antonin Chambolle\textsuperscript{c}, Peter Richtárik\textsuperscript{d,e,f}, Jonathan Schott\textsuperscript{g}, and Carola-Bibiane Schönlieb\textsuperscript{a}

\textsuperscript{a}Department for Applied Mathematics and Theoretical Physics, University of Cambridge, UK
\textsuperscript{b}Translational Imaging Group, CMIC, Department of Medical Physics & Biomedical Engineering, University College London, UK
\textsuperscript{c}CMAP, Ecole Polytechnique, CNRS, France
\textsuperscript{d}Visual Computing Center & Extreme Computing Research Center, KAUST, Thuwal, Saudi Arabia
\textsuperscript{e}School of Mathematics, University of Edinburgh, UK
\textsuperscript{f}The Alan Turing Institute, London, UK
\textsuperscript{g}Dementia Research Centre, Institute of Neurology, University College London, UK

ABSTRACT

Image reconstruction in positron emission tomography (PET) is computationally challenging due to Poisson noise, constraints and potentially non-smooth priors—let alone the sheer size of the problem. An algorithm that can cope well with the first three of the aforementioned challenges is the primal-dual hybrid gradient algorithm (PDHG) studied by Chambolle and Pock in 2011. However, PDHG updates all variables in parallel and is therefore computationally demanding on the large problem sizes encountered with modern PET scanners where the number of dual variables easily exceeds 100 million. In this work, we numerically study the usage of SPDHG—a stochastic extension of PDHG—but is still guaranteed to converge to a solution of the deterministic optimization problem with similar rates as PDHG. Numerical results on a clinical data set show that by introducing randomization into PDHG, similar results as the deterministic algorithm can be achieved using only around 10\% of operator evaluations. Thus, making significant progress towards the feasibility of sophisticated mathematical models in a clinical setting.

Keywords: positron emission tomography, PET, total variation, convex optimization, primal-dual algorithms, saddle point problems, stochastic optimization, inverse problems

1. INTRODUCTION

Positron emission tomography (PET) is an important medical imaging modality in many medical fields such as oncology and neurology as it allows monitoring of body functions.\textsuperscript{1,2} PET image reconstruction can be formulated as a convex optimization problem which is computationally challenging due the optimization set-up that includes Poisson noise, constraints and potentially non-smooth priors. An algorithm that can cope well with the aforementioned challenges is the primal-dual hybrid gradient algorithm (PDHG)\textsuperscript{3–6} which can decompose the problem into a sequence of easy operations such as matrix-vector products and proximal operators that have closed-form solutions. However, as the PDHG updates are parallel, every iteration of PDHG is computationally demanding for the large variable sizes (easily exceeding 100 million) encountered with modern PET scanners.
In this work, we numerically study the usage of SPDHG\textsuperscript{7}—a stochastic extension of the PDHG that makes use of the dual separability of the problem (cf. e.g.\textsuperscript{8–11}). Instead of updating all dual variables in parallel, in each iteration only a randomly selected subset of the dual variables are updated. Despite its randomness, it is still guaranteed to converge to a solution of the deterministic optimization problem with similar rates as PDHG.\textsuperscript{7} Conceptually, the randomness in the dual update is similar to the Kacmarz algorithm\textsuperscript{12–15} (also called ART\textsuperscript{16}), stochastic gradient descent and surrogate approaches with subsets.\textsuperscript{17,18} Unlike other methods,\textsuperscript{19–21} SPDHG does not rely on any consistency conditions on the system of linear equations nor a “good” subset selection for convergence. This robustness comes from the fact that—as in incremental optimization algorithms\textsuperscript{22–26}—we progressively build up information from previous iterations which reduces variance. An interesting feature of SPDHG is that it can be run with any iteration-independent sampling.

Over the course of the last years, also other random variants of PDHG have been proposed,\textsuperscript{8–10,27–32} some being special cases of SPDHG.\textsuperscript{8–10} A few comparisons have been made on small simulated data.\textsuperscript{7} An extensive comparison of these methods is out of the scope of this contribution.

2. MATHEMATICAL MODEL FOR PET RECONSTRUCTION

In this section we briefly describe the optimization model used for PET reconstruction.

Before the actual PET scan is performed, a radioactive tracer is injected into the patient.\textsuperscript{1,2} The radioactive tracer emits positrons which in turn annihilate with free electrons in the human tissue resulting in two photons traveling in (almost) opposite directions and measured by rings of detectors around the patient. Assuming the directions to be exactly opposite, the detected photons yield information about the line integral of the tracer along the line that connects the detectors that recorded these photons. The reconstruction task in PET is to estimate the distribution of the tracer given more than a billion of these events (example taken from a data set acquired on the Siemens Biograph mMR).

Mathematically, the reconstruction task is usually posed as the maximum a-posteriori estimation\textsuperscript{1,2} of an image given the measured data which is equivalent to the minimization problem

$$\min_{x \in \mathbb{X}} \left\{ \sum_{i=1}^{N} \varphi_i(a_i^T x) + \psi(x) \right\}$$

(1)

where for each $i \in \{1, \ldots, N\}$, the function $\varphi_i : \mathbb{R} \to \mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ (also called Kullback–Leibler divergence)

$$\varphi_i(y) = \begin{cases} y + r_i - b_i + b_i \log \left( \frac{b_i}{y + r_i} \right) & \text{if } y + r_i > 0 \\ \infty & \text{else} \end{cases}$$

(2)
measures the deviance from the estimated line integral $a_i^T x + r_i$ to the measured data $b_i \in \mathbb{N}$ in a Poisson sense where it is convention to define $0 \log 0 := 0$. The functions $\varphi_i$ are a distance measure, in the sense that $\varphi_i(y) \geq 0$ and $\varphi_i(y) = 0$ if and only if $a_i^T y + r_i = b_i$. Note that the $\varphi_i$ are convex but not strictly or strongly convex and not differentiable (on the whole domain $\mathbb{R}$).

The vectors $a_i = \eta_i w_i$ do not only contain the vector of coefficients for the line integral between the $i$th detector pair $w_i$, but also contain weights $\eta_i > 0$ to compensate for attenuation and detector efficiency. The background $r_i > 0$ is estimated from the data prior to reconstruction to compensate for non-linear effects such as scatter and randoms. Figure 1 shows parts of the data prior to reconstruction, together with the factors $\eta_i$ and the background $r_i$.

The space of all possible (discrete) tracer distributions is denoted by $X = \mathbb{R}^d := \mathbb{R}^{d_1 \times d_2 \times d_3}$. The function $\psi : X \rightarrow \mathbb{R}_\infty$ models the a-priori knowledge about the desired solution such as constraints or regularity. It is standard to assume the tracer is non-negative. Moreover, we model the tracer’s regularity with the total variation.\textsuperscript{33,34} Thus, the prior function $\psi$ becomes

$$\psi(x) = \lambda \|\nabla x\|_{1,2} + \nu_{\geq 0}(x) \quad (3)$$

where $\lambda > 0$ denotes the regularization parameter, and the gradient operator $\nabla x \in X^3$ discretized by forward differences, cf. e.g.\textsuperscript{5} for details. The $1,2$-norm of these gradients is defined as $\|y\|_{1,2} := \sum_\alpha \sqrt{\sum_{i=1}^3 (\nabla_i y_\alpha)^2}$ and the constraint function as

$$\nu_{\geq 0}(x) := \begin{cases} 0 & \text{if } x \geq 0 \\ \infty & \text{else} \end{cases} \quad (4)$$

For computational efficiency, it is usually desirable to reformulate problem (1) and to group measurements together into so-called “subsets” $B_i \subset \{1, \ldots, N\}, i = 1, \ldots, n - 1$ that are mutually disjoint $B_i \cap B_j = \emptyset$ if $i \neq j$ and cover all measurements $\bigcup_{i=1}^{n-1} B_i = \{1, \ldots, N\}$. Two examples of subset selection is shown in Fig. 2. While “subsets” is a medical imaging terminology, this is often being referred to as “mini-batches” in machine learning. We then define the measurement operator for a subset $i$ as $A_i := [a_j]_{j \in B_i}$, where the vectors $a_j$ form the rows of $A_i$ and the data fit as

$$f_i(A_i x) := \sum_{j \in B_i} \varphi_i(a_j^T x) \quad (5)$$

With $f_n(y) := \lambda \|y\|_{1,2}$, $A_n := \nabla$, $g(x) := \nu_{\geq 0}(x)$, we can rewrite (1) into the standard form

$$\min_{x \in X} \left\{ \sum_{i=1}^n f_i(A_i x) + g(x) \right\} \quad (6)$$
3. STOCHASTIC PRIMAL-DUAL HYBRID GRADIENT METHOD

The following exposition holds true in general, and not only for the special case of PET image reconstruction. We only assume that the $f_i$, $i = 1, \ldots, n$ and $g$ are proper, convex, and and lower semi-continuous. Under those assumptions,

$$f_i(z_i) = f_i^*(z_i) = \sup_{y_i} \langle z_i, y_i \rangle - f_i^*(y_i)$$  \hspace{1cm} (7)

where $f_i^*$ is the Fenchel conjugate of $f_i$ (and $f_i^{**}$ its biconjugate defined as the conjugate of the conjugate).\textsuperscript{35} Thus, instead of solving the optimization problem (6) directly, it can be reformulated as a saddle point problem

$$\min_x \sup_{y_1, \ldots, y_n} \left\{ \sum_{i=1}^n \langle A_i x, y_i \rangle - f_i^*(y_i) + g(x) \right\} . \hspace{1cm} (8)$$

A very popular algorithm to solve saddle point problems of the form (8) is the primal-dual hybrid gradient (PDHG)\textsuperscript{3–6} which reads (with extrapolation on the dual variable $y_i$)

$$x^{k+1} = \text{prox}_{\tau g} \left( x^k - \tau \sum_{i=1}^n A_i^T y_i^k \right) \hspace{1cm} (9)$$

$$y_i^{k+1} = \text{prox}_{\sigma f_i} \left( y_i^k + \sigma A_i x^{k+1} \right) \hspace{1cm} i = 1, \ldots, n \hspace{1cm} (10)$$

$$y_i^{k+1} = y_i^{k+1} + \theta (y_i^{k+1} - y_i^k) \hspace{1cm} i = 1, \ldots, n \hspace{1cm} (11)$$

It consists of simple operations like matrix-vector product and seemingly complicated operations like evaluations of\textit{proximal operators} which are defined as

$$\text{prox}_{\tau f}(y) := \arg \min_x \left\{ \frac{1}{2} \|x - y\|^2 + \tau f(x) \right\} . \hspace{1cm} (12)$$

However, for many problems—including PET reconstruction as established in the previous section—these proximal operators can be solved in closed-form and are separable,\textsuperscript{5–7} thus, allowing for efficient parallel implementation. The PDHG is guaranteed to converge if the step size parameters $\sigma, \tau$ are positive and satisfy $\sigma \tau \|A\|^2 < 1$ and $\theta = 1.\textsuperscript{5,6}$

While the PDHG is computationally attractive as it only needs simple operations, in every iteration many of those have to be executed. For instance, every iteration needs the evaluation of all operators $A_i$ and their adjoints $A_i$ for $i = 1, \ldots, n$. To overcome this issue, a randomized variant of PDHG has recently been introduced\textsuperscript{7} depicted in Algorithm 1. In each iteration, instead of all $n$ variables $y_i$, only one of those is updated. The sampling of those variables is allowed to be arbitrary as long as the probability $p_i$—that is selected during each iteration—is positive and constant during the iterations. Such a sampling is called “serial” and “proper”.\textsuperscript{5,6,36,37} Other samplings are possible, too, and convergence for a much wider class can be proven,\textsuperscript{7} but for simplicity we stick to this special case. Note that SDDH not only allows for a randomized update of the dual variables, but also allows the dual step size $\sigma_i$ to be different for each variable $y_i$. Moreover, if $n = 1$ and the dual variable gets updated in each iteration, i.e. $p_1 = 1$, then SPDHG coincides with the deterministic PDHG. The following theorem establishes convergence of SPDHG in terms of the generalized Bregman distance

$$D_\mathbf{g}^\mathbf{t}(x, x^\mathbf{t}) := g(x) - g(x^\mathbf{t}) - \langle \mathbf{g}^\mathbf{t}, x - x^\mathbf{t} \rangle \hspace{1cm} (13)$$

where $\mathbf{g}^\mathbf{t} \in \partial g(x^\mathbf{t})$ is an element of the subgradient of $g$ at $x^\mathbf{t}$. 
Algorithm 1 Stochastic Primal-Dual Hybrid Gradient algorithm (SPDHG). Input: primal and dual variable \(x^0, y^0\), step length parameters \(\sigma_i, \tau, \theta\), probabilities \(p_i\). Initialize: \(z^0 = A^*y^0, \bar{z}^0 = z^0\)

for \(k = 0, 1, \ldots \) do
\[
x^{k+1} = \text{prox}_{\tau y} (x^k - \tau \bar{z}^k)
\]
Select \(j \in \{1, \ldots, n\}\) with probability \(p_j > 0\).
\[
y^{k+1}_i = \begin{cases} \text{prox}_{\sigma_i f^i} \left(y^k_i + \sigma_i A_i x^{k+1}_i\right) & \text{if } i = j \\ y^k_i & \text{else} \end{cases}
\]
\[
\Delta z^{k+1}_i = A^*_j (y^{k+1}_j - y^k_j), \quad z^{k+1} = z^k + \Delta z^{k+1}, \quad \bar{z}^{k+1} = \bar{z}^k + \frac{\theta}{p_j} \Delta z^{k+1}
\]
end for

Theorem 1 (Convergence of SPDHG\(^7\)). Let \(\gamma < 1\), the extrapolation parameter \(\theta = 1\) and the step sizes \(\tau, \sigma_i\) be chosen such that
\[
\frac{\sigma_i \tau \|A_i\|^2}{p_i} \leq \gamma^2 \quad i = 1, \ldots, n.
\]
Then, the iterates \((x^k, y^k)\) of SPDHG (Alg. 1) almost surely converge to a saddle point \((x^s, y^s)\) in a Bregman sense:
\[
D_g^s (x^k, x^s) \to 0, \quad D_f^s (y^k, y^s) \to 0 \quad \text{a.s.}
\]
Moreover, the ergodic sequence \((x_K, y_K) := \frac{1}{K} \sum_{k=1}^K (x^k, y^k)\) converges to a saddle point with rate \(1/K\) in an expected Bregman sense:
\[
\mathbb{E} \left\{ D_g^s (x_K, x^s) + D_f^s (y_K, y^s) \right\} \leq \frac{c}{K}.
\]
For an explicit formula of the constant \(c\), see.\(^7\)

Remark. Stronger convergence rates—in line with the deterministic results\(^5,6\)—can be proven\(^7\) if more regularity is assumed on \(g\) and/or \(f_i\).

4. NUMERICAL RESULTS

Software The numerical examples are implemented in python using numpy and the operator discretization library (ODL).\(^38\) The ray tracing for the PET forward operator is performed using CUDA-based NiftyPET.\(^39\)

Data The clinical data are from a 70 year old cognitive normal subject, scanned as part of the “Insight 46”—a neuroscience sub-study of the MRC National Survey of Health and Development.\(^40\) The PET data is accompanied by the T1-, T2-weighted MRI, all acquired simultaneously on a Siemens Biograph mMR PET-MRI scanner. The PET data was acquired dynamically for 60 mins using the amyloid tracer 18F-florbetapir tracer with activity of 409 MBq at the time of injection. The subject’s -map was synthesized from the T1- and T2-weighted images for better accuracy.\(^41\) The MRI was co-registered to the PET image using NiftyReg software.\(^42\) Randoms and scatter have been computed prior to reconstruction.\(^39,43,44\)

Sampling The sampling is chosen to be non-uniform to allow for the difference between the \(n - 1\) subsets that correspond to the data and the subset that corresponds to the gradient. In each iteration we uniformly choose either data or gradient. In case the data is selected, then we again choose uniformly among the data subsets. Thus, overall the probability that subset \(i\) is selected is given by
\[
p_i = \begin{cases} \frac{1}{n} & \text{if } i < n \\ \frac{1}{\pi} & \text{if } i = n \end{cases}.
\]
After \(2(n - 1)\) iterations the SPDHG is expected to have evaluated all operators \(A_i\) once, thus, having computed the same number of rays as the deterministic PDHG after only one iteration.
Figure 3. PET reconstruction with TV prior; results after 100 expected operator evaluations. While PDHG still has artifacts in the reconstructed image, SPDHG is visually already very close to the saddle point.

Parameters In this experiment we choose $\gamma = 0.99$, $\theta = 1$. The number of data subsets varied between 1 (deterministic case), 21 and 252. The other step size parameters are chosen for PDHG as

$$\sigma_1 = \frac{\gamma}{\|A\|}, \quad \tau = \frac{\gamma}{\|A\|}$$

(18)

and for SPDHG as

$$\sigma_i = \frac{\gamma}{\|A_i\|}, \quad i = 1, \ldots, n, \quad \tau = \min_i \frac{\gamma p_i}{\|A_i\|}.$$  

(19)

The norms of the operators are estimated with 100 power iterations and allowing for an error of 5%.

Results Some example reconstructed images are found in Figs. 3 and 4 quantitative results in Fig. 5. The saddle point for the visual and quantitative comparisons has been approximated with 2000 iterations of the deterministic PDHG. It can be seen from the reconstructed images after 100 expected operator evaluations in 3 that SPDHG is much faster than the deterministic PDHG. While the reconstruction with PDHG still has artifacts, the images that are reconstructed with SPDHG is visually very similar to the saddle point already after 100 iterations. A comparison of SPDHG with 21 and 252 subsets with PDHG after only 20 expected operator evaluations. The result of PDHG is clinically useless, as major anatomical structures are not yet visible. On the other hand, with more and more subsets, even with this little effort, a reasonable (although not perfect) image can be reconstructed. It is also easy to see, that the artifacts get reduced with more subsets indicating that with more subsets the algorithm converges faster. This is confirmed by the quantitative results in 5, where the same trend is observable for all four figures of merit.
5. CONCLUSIONS AND FUTURE WORK

A bottleneck in the translation of mathematical ideas into clinical applications in PET imaging has been the computational effort that modern algorithms need in order to solve the optimization problems. In this work we have numerically evaluated SPDHG—a stochastic variant of PDHG—to reconstruct images from clinical data with a total variation prior. As the results clearly indicated, by introducing randomness into the algorithm a similar performance as the deterministic variant may be achieved with only around 10% of operator evaluations and thereby cutting the computational effort significantly. The algorithm is general enough to solve the optimization problem with other priors such as the directional total variation / parallel level sets,\textsuperscript{45,46} total generalized variation (TGV),\textsuperscript{47-49} directional TGV\textsuperscript{50} or structure tensor-based total variation\textsuperscript{51} to name a few. In addition to PET reconstruction, SPDHG may help easing the computational effort in multi-modal medical imaging as well, where the PDHG has been used to jointly reconstruct from simultaneously acquired PET-MRI data with a nuclear norm variant of TGV\textsuperscript{52} or coupled Bregman iterations.\textsuperscript{53} Similarly, a comparable speed up is expected in other imaging modalities that involve line integrals like computerized tomography (CT) where PDHG also found its applications.\textsuperscript{54}

Future work will encompass further numerical investigations of PET image reconstruction with different priors, other sampling strategies and step-size choices. While the randomness already allowed a significant speed up, we will investigate the use of smoothness and strong-convexity for algorithmic acceleration to increase the speed-up even further.

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 Math-
Figure 5. Quantitative evaluation of reconstruction algorithms. In all four measures, both variants of SPDHG are much faster than the deterministic PDHG. Moreover, with more subsets the algorithm becomes even faster.

REFERENCES


