Stochastic EM methods with variance reduction for penalised PET reconstructions

To cite this article: Željko Kereta *et al* 2021 *Inverse Problems* **37** 115006

View the article online for updates and enhancements.

Stochastic EM methods with variance reduction for penalised PET reconstructions

Željko Kereta[1,*] ⓘ **, Robert Twyman**[2] **, Simon Arridge**[1] ⓘ **,
Kris Thielemans**[2] ⓘ **and Bangti Jin**[1] ⓘ

[1] Department of Computer Sciences, University College London, Gower Street, London WC1E 6BT, United Kingdom
[2] Institute of Nuclear Medicine, University College London, London, United Kingdom

E-mail: z.kereta@ucl.ac.uk, s.arridge@ucl.ac.uk, b.jin@ucl.ac.uk, robert.twyman.18@ucl.ac.uk and k.thielemans@ucl.ac.uk

CrossMark

Abstract

Expectation-maximisation (EM) is a popular and well-established method for image reconstruction in positron emission tomography (PET) but it often suffers from slow convergence. Ordered subset EM (OSEM) is an effective reconstruction algorithm that provides significant acceleration during initial iterations, but it has been observed to enter a limit cycle. In this work, we investigate two classes of algorithms for accelerating OSEM based on variance reduction for penalised PET reconstructions. The first is a stochastic variance reduced EM algorithm, termed as SVREM, an extension of the classical EM to the stochastic context that combines classical OSEM with variance reduction techniques for gradient descent. The second views OSEM as a preconditioned stochastic gradient ascent, and applies variance reduction techniques, i.e., SAGA and SVRG, to estimate the update direction. We present several numerical experiments to illustrate the efficiency and accuracy of the approaches. The numerical results show that these approaches significantly outperform existing OSEM type methods for penalised PET reconstructions, and hold great potential.

Keywords: positron emission tomography, stochastic gradient, variance reduction, expectation maximization, ordered subset expectation maximisation

---

*Author to whom any correspondence should be addressed.

(Some figures may appear in colour only in the online journal)

1. Introduction

Positron emission tomography (PET) is a nuclear imaging technique that allows for the measurement of biochemical changes in the body by observing the spatial distribution of a radioactive tracer. Positron emitting radionuclides are attached to a biochemical compound to create a radioactive tracer, e.g. fluorodeoxyglucose, that is used in natural metabolic processes by an organ or tissue of interest. The radionuclides decay and the emitted positron travels a short distance before encountering and annihilating with an electron. This annihilation interaction results in a pair of 511 keV photons that travel anti-parallel. The emitted photons may be measured by a pair of detector elements along a ring of crystalline detectors surrounding the subject. If two photons are detected within a short coincidence timing window, a PET scanner will record a coincidence event along the line-of-response between the two measuring detectors. The goal of PET image reconstruction problem is then to reconstruct an estimate of the emission distribution from the measured coincidence data. This inverse problem is ill-posed in the sense of Hadamard since the solution to the problem is not stable with respect to the perturbation in the data.

Iterative methods have been widely used in PET reconstruction [34], amongst which the expectation maximisation (EM) algorithm and its various variants, e.g., MLEM [16], OSEM [24], RAMLA [3], BSREM [1, 13] and OS-SPS [18], are predominant. Shepp and Vardi [37, 38] reformulated the PET reconstruction problem into a maximum likelihood (ML) estimation of the tracer distribution, and developed an iterative scheme via the EM algorithm (MLEM), which enjoys several desirable features, e.g., a closed-form for iterate updates and nonnegativity preservation. The EM algorithm consists of two steps: (i) the E-step computes the complete data sufficient statistic; and (ii) the M-step updates the estimate by maximising the complete data log-likelihood. The algorithm converges monotonically (in decreasing the objective) but slowly. Moreover, full batch updates (i.e. using all measured data to compute the sufficient statistic) can be costly for large data sizes. To mitigate the ill-posedness of the PET reconstruction problem, a suitable penalty is employed, leading to a maximum *a posteriori* (MAP) problem [23, 25]. This requires adapting the standard EM algorithm, since a closed-form solution at the M-step is often no longer available [34, p R561]. There are several approaches to address the challenge, such as the one-step-late algorithm [21] (applicable to differentiable penalty terms but generally not convergent to the MAP solution), or more principled methods via modified EM algorithms [12] or separable parabolic surrogates for the penalty [18].

One established procedure to mitigate the aforementioned computational challenge with full batch data is ordered subset EM (OSEM) [24], which first divides the measured data into disjoint subsets and then applies the EM algorithm to one subset at each iteration, either in a cyclic or a stochastic manner [24]. This greatly reduces the cost per update, and leads to significant acceleration during initial iterations. However, the standard OSEM algorithm has been observed to often not converge but instead enters a limit cycle [4, 13]. This has motivated intensive development of modified OSEM algorithms that retain both the speed-up in early iterations but also exhibit convergence to the MAP solution (e.g., by suitably adjusting the step-size schedule) [1, 3, 13].

EM and OSEM can also be written in a gradient ascent-like form, where the search direction is a preconditioned gradient of the objective [28]. This viewpoint enables designing a range of new methods that allow a general class of differentiable penalties. This idea has

recently been experimentally studied for PET with iteration-dependent and constant precon-ditioners in [41, 42], respectively, both for standard stochastic ascent approaches and for variance reduction methods. For nonsmooth convex penalties, e.g., total (generalised) vari-ation [2], gradient approaches are no longer directly applicable, and one may resort to a saddle point reformulation, and then update primal and dual variables accordingly [7]. Chambolle *et al* [6, 17] developed a stochastic variant of such an algorithm using only one random component of the dual variable at each update, and provide a convergence guarantee. Alternatively, one may employ proximal methods to handle non-smooth penalties, and many variance reduction algorithms have been extended to the proximal setting [44].

In this work, we contribute to stochastic variance reduction algorithms for the MAP problem in PET, for a popular class of penalty terms, by drawing on recent advances in stochastic optimi-sation and machine learning. First, we develop a novel algorithm, termed as stochastic variance reduced EM (SVREM), for the MAP reconstruction. It is motivated by the Online-EM [5] (see also [31]) and its variance reduction variants [9, 27], originally developed for an un-penalised problem. We present an extension to the MAP problem by combining variance reduced EM for computing a variance reduced running average of the sufficient statistics with the surrogate approach for the penalty [8]. The resulting SVREM algorithm maintains the EM nature for PET reconstruction, e.g., nonnegativity preservation, and admits an explicit maximiser at each M-step. The overall algorithm is mathematically principled and numerically easy to implement. To summarise, SVREM extends the framework in [9] to the penalised PET reconstruction problem by incorporating the surrogate approach for the penalty. Second, we revisit variance reduction algorithms for stochastic gradient ascent (SGA), and their use in iterative PET reconstruction, which were recently experimentally studied [41, 42]. These algorithms do not belong to the EM family, but rather to the class of diagonally preconditioned gradient ascent algorithms. Due to the inclusion of the penalty, the non-negativity of the iterates is no longer ensured, which requires a projection step (i.e., the proximal map of the characteristic function on the admissible set $\{f \geqslant 0\}$). In theorem 4.1, we show the almost sure convergence to a maximiser for a mod-ified likelihood with a constant preconditioner. Third and last, we conduct extensive numerical experiments, which show that these algorithms enjoy steady convergence, outperform classical OSEM type methods, and are very promising for penalised PET reconstruction.

The rest of the paper is organised as follows. In section 2, we describe the mathematical formulation of the ML problem in PET, and EM and its stochastic variants for ML estimation. Then in section 3, we discuss the EM algorithm for MAP reconstruction using parabolic surro-gates. In section 4, we discuss a second class of numerical algorithms, i.e., variance reduction algorithms based on gradient ascent for the MAP problem. Last, in section 5 we present numer-ical results that examine and illustrate features of these algorithms. Throughout, the notation $\stackrel{c}{=}$ denotes expressions that are equal up to an additive constant that is independent of a function's argument. The notation $\oslash$ and $\odot$ denote entrywise division and multiplication of matrices or vectors. For any fixed $M \in \mathbb{N}$ we denote $[M]$ to be the set $[M] = \{1, \ldots, M\}$. We use lower-case letters, e.g., g and f for column vectors, and upper case letters for matrices and operators. For a matrix A we let a_m denote its mth row. The notation $P_{\geqslant 0}(x)$ denotes the coordinate-wise projection of x onto the non-negative plane, i.e., $P_{\geqslant 0}(x) = (\max(0, x_n))_{n=1}^N$. The notation 1 is slightly abused for a constant vector of suitable size with all entries equal to one.

2. Expectation maximisation and its stochastic variants

In this section, we describe the ML PET problem, and the EM algorithm and its stochastic variants for finding ML solutions.

2.1. ML PET problem

First we recall the standard mathematical formulation of the PET reconstruction problem. Let M denote the number of detector bins, and g_m the number of emissions detected in the mth bin, so that the measured data are $g = (g_1, \ldots, g_M)^\top \in \mathbb{R}^M$. It is customary to approximate the measurement means of the unknown tracer distribution in the form of a linear problem

$$\mathbb{E}[g] = Af + r,$$

where $A \in \mathbb{R}^{M \times N}$ is the system matrix with nonnegative elements, $f \in \mathbb{R}^N$ is vector of voxel values, and $r \geqslant 0$ represents the mean number of background events such as scatters, background radiation, and random coincidences, which will often be omitted below. Emission measurements in the mth bin are modelled by the following Poisson model:

$$g_m \sim \text{Poisson}(\mathbb{E}[g_m]).$$

Recall that a random variable g follows the Poisson distribution $\text{Poisson}(\lambda)$ with a parameter $\lambda > 0$ if $\text{Prob}(g = k) = \frac{\lambda^k e^{-\lambda}}{k!}$, for $k = 0, 1, \ldots$. Assuming that detector bins record independent measurements, and conditioning on the tracer distribution f, it follows that the probability distribution function $p(g|f)$ of the emission measurements g is given by

$$p(g|f) = \prod_{m=1}^{M} \exp(-\mathbb{E}[g_m]) \frac{\mathbb{E}[g_m]^{g_m}}{g_m!}. \tag{2.1}$$

The ML estimator f_{ml} of f is computed by maximising the likelihood $p(g|f)$ in (2.1), or equivalently its logarithm. Omitting terms independent of f, this yields the following objective

$$\mathcal{L}(f) := \log(p(g|f)) \overset{c}{=} \sum_{m=1}^{M} (-a_m^\top f - r_m + g_m \log(a_m^\top f + r_m)). \tag{2.2}$$

The term $g_m \log(a_m^\top f + r_m)$ is defined to be 0 when $g_m = 0$, for any value of the logarithm [1]. The ML estimator f_{ml} is then defined as

$$f_{\text{ml}} \in \underset{f \geqslant 0}{\text{argmax}} \mathcal{L}(f). \tag{2.3}$$

The functional $\mathcal{L}(f)$ is concave on the space of all admissible tracer distributions ($f \geqslant 0$), but a direct solution via Karush–Kuhn–Tucker conditions is intractable, and instead iterative approaches are commonly used, which we discuss in more detail in section 2.2 below.

Next we introduce the concept of ordered subsets. Consider a partition $\mathcal{S} = \{S_1, \ldots, S_{N_s}\}$ of the set $[M]$, i.e. a collection of (sub)sets such that $\emptyset \neq S_t \subset [M]$; $S_{t_1} \cap S_{t_2} = \emptyset$ for $t_1 \neq t_2$; and $\cup_{t=1}^{N_s} S_t = [M]$. For a vector v and a matrix A, we denote by v_t and A_t the subvector of length $|S_t|$ and an $|S_t| \times N$ submatrix whose entries, respectively row indices, belong to S_t. Then given the partition $\mathcal{S}$, we can subdivide the log likelihood $\mathcal{L}(f)$ into

$$\mathcal{L}(f) = \sum_{t=1}^{N_s} \mathcal{L}_t(f), \quad \text{with} \quad \mathcal{L}_t(f) = \sum_{m \in S_t} (-a_m^\top f - r_m + g_m \log(a_m^\top f + r_m)).$$

The partition $\mathcal{S}$ needs to be carefully constructed, in order to optimise the quality of the reconstructions [24]. Moreover, subsets should be balanced so that emission probabilities $\sum_{m \in S_t} a_{mn}$ are nearly independent of the subset index t [4], which is what we also adhere to. Thus, it is recommended that subset iterations should follow an order such that projections corresponding to the next subset are as 'perpendicular' as possible to previous ones [22].

2.2. ML expectation maximisation

EM is the most well-known example of an iterative, functional substitution scheme for PET reconstruction. It solves the ML problem (2.3) by replacing the objective (2.2) through a complete data framework. We follow the complete data framework due to [37]. Let $G \in \mathbb{R}^{M \times N}$ and $G_t \in \mathbb{R}^{|S_t| \times N}$ denote the full and subset complete data matrices, respectively, with entries g_{mn} that denote the number of emissions detected in bin m that originated from voxel site n. For notational simplicity, we omit the background term r_m below. Since

$$\mathbb{E}[g_{mn}|f] = a_{mn}f_n \quad \text{and} \quad \mathbb{E}[g_m] = \sum_{n=1}^{N} \mathbb{E}[g_{mn}],$$

the (subset) complete data likelihood $p(G_t|f)$ corresponding to the subset S_t satisfies

$$p(G_t|f) = \prod_{m \in S_t} \prod_{n=1}^{N} p(g_{mn}|f) = \prod_{m \in S_t} \prod_{n=1}^{N} e^{-\mathbb{E}[g_{mn}|f]} \frac{\mathbb{E}[g_{mn}|f]^{g_{mn}}}{g_{mn}!}.$$

Now consider the conditional expectation

$$\mathbb{E}_{G|g,f}[\log p(G|f)] = \sum_{t=1}^{N_s} \mathbb{E}_{G_t|g_t,f}[\log p(G_t|f)],$$

with

$$\mathbb{E}_{G_t|g_t,f}[\log p(G_t|f)] \overset{c}{=} \sum_{n=1}^{N} e_n^t(f)f_n \, \log(f_n) - f_n \sum_{m \in S_t} a_{mn}, \quad \text{and}$$

$$e_n^t(f) = \sum_{m \in S_t} \frac{a_{mn}g_m}{\sum_{l=1}^{N} a_{ml}f_l}. \tag{2.4}$$

Each iteration of the OSEM algorithm consists of two steps:

E step. For a given subset index t_k, compute the expectation

$$\mathbb{E}_{G_{t_k}|g_{t_k},f}[\log p(G_{t_k}|f)]; \tag{2.5}$$

M step. Maximise the expectation by

$$f^{k+1} = \underset{f \geqslant 0}{\operatorname{argmax}} \mathbb{E}_{G_{t_k}|g_{t_k},f}[\log p(G_{t_k}|f)] = \left(f_n^k e_n^{t_k}(f^k) / \sum_{m \in S_{t_k}} a_{mn} \right)_{n=1}^{N}. \tag{2.6}$$

When $N_s > 1$, the above algorithm is referred to as OSEM. The standard EM algorithm uses $N_s = 1$ subset and obeys the update rule

$$f^{k+1} = \left(f_n^k e_n(f^k) / \sum_{m=1}^{M} a_{mn} \right)_{n=1}^{N}, \quad \text{with} \quad e_n(f) = \sum_{m=1}^{M} \frac{a_{mn}g_m}{\sum_{l=1}^{N} a_{ml}f_l}.$$

Like before, the term $\frac{g_m}{\sum_{l=1}^{N} a_{ml} f_l}$ is set to 0 whenever the denominator is zero. Clearly, EM and OSEM preserve nonnegativity of the updates.

The above framework falls under the umbrella of function substitution-type methods [30], which at each step replace the original objective function with a surrogate. Recall that a function $\widehat{\Phi}$ is said to be a surrogate of a concave objective Φ if it satisfies

$$\Phi(f) - \Phi(f^k) \geqslant \widehat{\Phi}(f; f^k) - \widehat{\Phi}(f^k; f^k) \quad \text{and} \quad \nabla \Phi(f)|_{f=f^k} = \nabla \widehat{\Phi}(f; f^k)|_{f=f^k}.$$

These defining properties ensure that maximising the surrogate $\widehat{\Phi}$ monotonically increases the value of the objective Φ, thereby guaranteeing the convergence of the objective value. It can be shown that $\mathbb{E}_{G_t|g_t,f}[\log p(G_t|f)]$ is a surrogate for $\mathcal{L}_t(f)$ [11, 29, 30, 38].

2.3. Stochastic expectation maximisation

The EM algorithm represents a powerful and versatile approach for inference and estimation involving distributions whose complete data likelihood belongs to the exponential family

$$p(G|f) = c_G \, \exp(\zeta(f)^\top s(G) - \mathcal{U}(f)),$$

$$\mathcal{U}(f) = \sum_{m=1}^{M} a_m^\top f, \quad \zeta(f) = (\log(f_n))_{n=1}^{N}, \qquad s(G) = \left(\sum_{m=1}^{M} g_{mn} \right)_{n=1}^{N}.$$

We can then rewrite MLEM as

$$f^{k+1} = \underset{f \geqslant 0}{\operatorname{argmax}} \zeta(f)^\top s(f^k) - \mathcal{U}(f), \tag{2.7}$$

where

$$s(f^k) = \mathbb{E}_{G|g,f^k}[\log s(G)] = f \odot (\nabla \mathcal{L}(f^k) + A^\top 1)$$

is the full sufficient statistic. Physically, $s(G) \in \mathbb{R}^N$ represents the unknown emission quantities: the nth entry of the full sufficient statistics $s(G)$ is $\sum_{m=1}^{M} g_{mn}$, the total number of emissions from the nth voxel. We can then interpret the E step in (2.5) as computing either the full expected statistic or its subset variant. A direct way to randomise OSEM is through a random sampling of the subset. This can be achieved by resampling the subset at each iteration, or choosing only the subset index at random (for a fixed partition $\mathcal{S}$). We employ the latter strategy since in practice it shows superior performance [42].

There have been several recent proposals [5, 9, 27, 45] that randomise the classical EM algorithm [16] differently and show excellent performance on a range of problems, e.g. Gaussian mixtures, natural language processing and hidden Markov models. One notable class of these algorithms start from the expression (2.7) and instead of computing the full expected statistic $s(f^k)$, or the corresponding subset statistics $\tau_{t_k}(f^k)$ defined below, at each iteration use an exponentially running approximation $\widehat{s}^k$ of $s(f^k)$. In the M-step we then compute

$$\widehat{f}^k = \underset{f \geqslant 0}{\operatorname{argmax}} \zeta(f)^\top \widehat{s}^k - \mathcal{U}(f). \tag{2.8}$$

To compute the estimate $\widehat{s}^k$ of $s(f^k)$, the common practice is to view the full statistic as an average of N_s subset statistics

$$s(f) = \frac{1}{N_s} \sum_{t=1}^{N_s} \tau_t(f),$$

where t is a subset index, and with subset statistics τ_t given by

$$\tau_t(f) = N_s f \odot (\nabla \mathcal{L}_t(f) + A_t^\top 1), \quad t \in [N_s].$$

Both $s(G)$ and τ_t are of the size of f. Then at each iteration, we (randomly) select an index t to update the estimate $\widehat{s}^k$ of the full statistic $s(f^k)$.

This idea was first proposed to derive an online EM algorithm [5, 31] for handling streaming data, in order to approximate the conditional statistics by exponentially moving averages as the data streams in. The resulting updates are akin to subset gradient updates. Specifically, for an initial guess $\widehat{s}^0$ and a subobjective index t_k, it can be written as

$$\widehat{s}^{k+1} = (1 - \alpha_k)\widehat{s}^k + \alpha_k \tau_{t_k}(\widehat{f}^k_{\text{sem}}), \tag{2.9}$$

where $\{\alpha_k\}_k$ is a decaying stepsize schedule, and the index t_k is drawn uniformly at random. Using the estimator in (2.9), we can compute the corresponding iterates, denoted by $\widehat{f}^k_{\text{sem}}$, by maximising the objective in (2.8). This algorithm is termed as stochastic expectation maximisation (SEM) below. The convergence of the resulting sequence of iterates $\widehat{f}^k_{\text{sem}}$ is highly dependent on the variance of the estimated statistics $\tau_{t_k}(f^k)$ (compared with $s(f^k)$ of the standard MLEM algorithm). Thus, the convergence guarantee requires a decaying stepsize schedule.

To reduce the variance of the gradient estimate (for stochastic optimisation) and to allow a constant stepsize, several variance reduction techniques have been developed in the machine learning community, e.g., SAG [35], SAGA [15], and SVRG [26]; see [20] for an overview. These techniques reduce the variance of the gradient estimate by including an estimator of the full gradient in the search direction, which is updated either according to a predefined update schedule, or per-iteration. For Online-EM, they can be used to reduce the variance of the sufficient statistics estimate. A variant of Online-EM, inspired by SVRG, was developed in [9], which uses an anchor point f^{anc} and a full but infrequently updated estimate s^{anc} of $s(f)$ at f^{anc}:

$$\widehat{s}^{k+1} = (1 - \alpha)\widehat{s}^k + \alpha(\tau_{t_k}(\widehat{f}^k_{\text{svrem}}) - \tau_{t_k}(\widehat{f}^{\text{anc}}) + s^{\text{anc}}) \tag{2.10}$$

If $k \bmod \eta N_s = 0$, set $f^{\text{anc}} = f^k_{\text{svrem}}$ and update $s^{\text{anc}} = s(f^{\text{anc}})$.

Then in analogy to SEM, we use (2.10) to compute the corresponding iterates $\widehat{f}^k_{\text{svrem}}$ by maximising (2.8). The resulting algorithm is referred to as SVREM. The full anchored expectation s^{anc} and the anchor image estimate f^{anc}, are updated once every $\eta \in \mathbb{N}$ epochs, where one epoch refers to every N_s iterations. It is worth noting that for both Online-EM and its variance reduced variants, the resulting updates still follow the EM paradigm, where the M-step uses the computed running estimate $\widehat{s}(f^k)$ of the sufficient statistic $s(f^k)$.

Remark 2.1. Naturally, other variance reduction techniques can also be applied to improve SEM. For example, an algorithm based on SAGA reads [27]

$$\widehat{s}^{k+1} = (1 - \alpha)\widehat{s}^k + \alpha \left(\tau_{t_k}(f^k) - s_{t_k}^k + \frac{1}{N_\mathrm{s}} \sum_{t=1}^{N_\mathrm{s}} s_t^k \right)$$

draw $\tilde{t}_k \in [N_\mathrm{s}]$ and set $s_{\tilde{t}_k}^{k+1} = \tau_{\tilde{t}_k}(f^k)$, and for $t \neq \tilde{t}_k$ keep $s_t^{k+1} = s_t^k$.

We shall not examine these variants, and focus only on SVREM.

The above methodology naturally extends to the ordered subset setting for the PET problem (2.2). Indeed, in the presence of background events, we have

$$\tau_t(f) = \mathbb{E}_{G_t|g_t, f^k}[\log s(G_t)] = N_\mathrm{s} \sum_{m \in S_t} \left(\frac{g_m a_{mn} f_n}{\sum_{\ell=1}^N a_{m\ell} f_\ell + r_m} \right)_{n=1}^N$$

$$= N_\mathrm{s} f \odot A_t^\top (g_t \oslash (A_t f + r_t)).$$

For any estimator $\widehat{s}^k$ of the full statistic $s(f^k)$, for the PET ML problem (2.3), the maximisation step in (2.8) still admits a unique solution and can be computed as (2.6).

Now we briefly comment on the convergence of SVREM (2.10). The convergence result in [9] requires the subset statistics $\tau_t(f)$ to be Lipschitz continuous, which holds only for nonzero backgrounds r_t. This condition arises also for standard MLEM [1]. However, any realistic PET scan indeed has nonzero background. Regardless, there are two common remedies, a practical and a theoretical one. The former is to set the pixel value to 0 whenever the denominator in (2.4) or (2.6) is equal to zero. The latter is to modify the likelihood term, using a quadratic approximation near the origin [1]. Specifically, let $\varphi_m(\ell) = g_m \log(\ell) - \ell$, and define

$$\hat{\varphi}_m(\ell) = \begin{cases} \dfrac{\varphi_m''(\varepsilon)}{2}(\ell - \varepsilon)^2 + \varphi_m'(\varepsilon)(\ell - \varepsilon) + \varphi_m(\varepsilon), & \text{if } \ell \leqslant \varepsilon \text{ and } r_m = 0 \\ \varphi_m(\ell), & \text{otherwise,} \end{cases}$$

(2.11)

with a constant $\varepsilon > 0$. If ε is sufficiently small, then the solution set does not change [1].

3. Stochastic variance reduced EM (SVREM) for penalised PET reconstruction

To address the inherent ill-posed nature of the ML problem (2.3), one popular approach is variational regularisation, which introduces a convex penalty $\mathcal{R}(f)$ [25]. This can often be interpreted as a MAP estimation and the corresponding estimator f_map is given by

$$f_\mathrm{map} = \underset{f \geqslant 0}{\mathrm{argmax}}\{\Phi(f) := \mathcal{L}(f) - \beta \mathcal{R}(f)\}. \tag{3.1}$$

A common type of penalties used in PET reconstruction take the form

$$\mathcal{R}(f) = \frac{1}{2} \sum_{n=1}^N \sum_{j \in \mathcal{N}_n} w_{nj} \rho \left(f_n - f_j \right),$$

where $w_{nj} \geqslant 0$ are weights, $\mathcal{N}_n$ is the neighbourhood of the nth voxel, and ρ is a potential function. The penalty $\mathcal{R}(f)$ is used to promote the desired image structure, which are often meant to

Table 1. Commonly used penalty terms.

	$\rho(f)$	$\rho'(f)$	$\gamma_\rho(f)$	Comments
Quadratic	$\frac{f^2}{2}$	f	1	Not edge preserving
Huber	$\begin{cases} \frac{f^2}{2}, & \|f\| \leqslant \delta \\ \delta\|f\| - \frac{\delta^2}{2}, & \|f\| \geqslant \delta \end{cases}$	$\begin{cases} f, & \|f\| \leqslant \delta \\ \delta, & \|f\| \geqslant \delta \end{cases}$	$\begin{cases} 1, & \|f\| \leqslant \delta \\ \frac{\delta}{\|f\|}, & \|f\| \geqslant \delta \end{cases}$	Not strictly convex nor $\mathcal{C}^2$
log cosh	$\delta^2 \log \cosh(f/\delta)$	$\delta \tanh(f/\delta)$	$\delta\frac{\tanh(f/\delta)}{f}$	
Hyperbola	$\delta\left(\sqrt{1+(f/\delta)^2} - 1\right)$	$\frac{f}{\sqrt{1+(f/\delta)^2}}$	$\frac{1}{\sqrt{1+(f/\delta)^2}}$	Approximates TV

be locally smooth but still preserve edge phenomena. Thus, ρ should be smooth within a given tissue or organ, while retaining sharp boundaries between different tissues. Most potentials are thus monotonic, nondecreasing functions of the difference $|f_n - f_j|$ that are roughly quadratic near the origin and linear away from the origin, and satisfy the following assumption.

Assumption 3.1. The potential $\rho(f)$ is symmetric, continuously differentiable, with $\rho'(f)$ nondecreasing (so that $\rho(f)$ is convex). The curvature function $\gamma_\rho(f) = \frac{\rho'(f)}{f}$ is assumed to be nonincreasing for $f \geqslant 0$, and such that $\lim_{f \searrow 0} \gamma_\rho(f)$ is finite and nonzero.

A list of commonly used penalty terms satisfying these properties is given in table 1. Note that this does not cover the relative difference penalty [32]. For the Huber, log cosh and hyperbola penalties, the parameter $\delta > 0$ controls the transition between the quadratic (smooth) and linear (edge-preserving) regimes of the given penalty term.

The general principle for solving the corresponding MAP problem (3.1) by EM does not change. That is, for MLEM, OSEM, SVREM, and SEM, instead of (2.8), we compute

$$\widehat{f}^k = \underset{f \geqslant \mathbf{0}}{\operatorname{argmax}} \zeta(f)^\top \widehat{s}^k - \mathcal{U}(f) - \beta \mathcal{R}(f). \tag{3.2}$$

If the penalty $\mathcal{R}(f)$ is separable (i.e., no coupling between the entries, which for example is the case for the quadratic prior), the objective function in (3.2) is separable and the M-step has a closed-form solution. However, this is not the case for many penalties of interest in PET reconstruction and computing the maximiser requires solving a coupled system of equations. Thus, (3.2) is often maximised iteratively [34].

To explicitly solve the M-step, we employ a separable surrogate of the penalty $\mathcal{R}(f)$. Surrogates have been widely used in penalised PET reconstruction [12, 13, 18, 19]. The idea is to construct a surrogate for the potential ρ, either to facilitate the computation of the prior or to improve conditioning (and convergence). We employ the parabolic surrogate [8]. Namely, consider the surrogate for the potential ρ given by

$$\widehat{\rho}^k(f_n; f_j) = \rho(f_n^k - f_j^k) + \rho'(f_n^k - f_j^k)\left(f_n - f_j - (f_n^k - f_j^k)\right)$$
$$+ \gamma_\rho(f_n^k - f_j^k)\left((f_n - f_n^k)^2 + (f_j - f_j^k)^2\right)$$
$$\overset{\mathsf{c}}{=} \gamma_\rho(f_n^k - f_j^k)\left(\left(f_n - \frac{f_n^k + f_j^k}{2}\right)^2 + \left(f_j - \frac{f_n^k + f_j^k}{2}\right)^2\right),$$

and define the surrogate penalty by

$$\widehat{\mathcal{R}}(f; f^k) = \frac{1}{2}\sum_{n=1}^{N} \sum_{j \in \mathcal{N}_n} w_{nj} \widehat{\rho}^k(f_n; f_j).$$

Then for $n \neq j$ the nth and the jth entry are decoupled since the partial derivatives $\frac{\partial \widehat{\rho}^k(f_n; f_j)}{\partial f_n}$ are given by

$$\frac{\partial \widehat{\rho}^k(f_n; f_j)}{\partial f_n} = \gamma_\rho(f_n^k - f_j^k)(2f_n - f_n^k - f_j^k).$$

The nth partial derivative for the surrogate objective $\zeta(f)^\top \widehat{s}^k - \mathcal{U}(f) - \beta\widehat{\mathcal{R}}(f; f^k)$ is given by

$$\frac{1}{f_n}\widehat{s}_n^k - 2\beta f_n \sum_{j \in \mathcal{N}_n} d_{nj} + \left(\beta f_n^k \sum_{j \in \mathcal{N}_n} d_{nj} + \beta \sum_{j \in \mathcal{N}_n} d_{nj}f_j^k - \sum_{m=1}^{M} a_{mn} \right), \qquad (3.3)$$

where $\widehat{s}^k$ is an estimator of the expected statistic, and $d_{nj} := w_{nj}\gamma_\rho(f_n^k - f_j^k)$. Equating (3.3) with zero gives a scalar equation of the form

$$af^{-1} - 2bf + c = 0,$$

with

$$a = \widehat{s}_n^k, \qquad b = \beta \sum_{j \in \mathcal{N}_n} d_{nj}, \qquad c = \beta f_n^k \sum_{j \in \mathcal{N}_n} d_{nj} + \beta \sum_{j \in \mathcal{N}_n} d_{nj}f_j^k - \sum_{m=1}^{M} a_{mn}.$$
$$(3.4)$$

Thus, we arrive at a quadratic equation, with a unique nonnegative solution which can be easily evaluated at each iteration. Provided that $b > 0$ it is given as

$$f = (4b)^{-1}(c + \sqrt{c^2 + 8ab}), \qquad (3.5)$$

where we note that the discriminant is nonnegative under assumption 3.1, and when $b = 0$, i.e. when there is no prior, (3.3) is linear, and the unique solution is given as $f = -c/a$. SEM and SVREM for penalised PET reconstruction then proceed as follows. At each iteration $k \in \mathbb{N}$, we first select the subset index t_k. We then update the estimator of the conditional statistic according to (2.9) and (2.10), respectively. Coefficients in (3.4) are then evaluated and we compute the corresponding iterates through (3.5).

4. Stochastic EM algorithm based on gradient ascent

In this section, we describe a second class of algorithms for problem (3.1). It is inspired by the following additive formulation of the EM update (2.6) [28]:

$$f_{\text{osem}}^{k+1} = f_{\text{osem}}^k + (f_{\text{osem}}^k \oslash A_{t_k}^\top 1) \odot \nabla\mathcal{L}_{t_k}(f_{\text{osem}}^k), \quad \text{with} \quad \nabla\mathcal{L}_t(f) = A_t^\top(g_t \oslash A_tf - 1).$$
$$(4.1)$$

This can be interpreted as a preconditioned gradient ascent. It is then natural to replace the gradient $\nabla\mathcal{L}(f)$ of the likelihood $\mathcal{L}(f)$ with that of $\Phi(f)$. This strategy is directly amenable to stochastic gradients, which are very appealing due to their low cost per iteration, flexibility with the penalty (i.e., there is no need for constructing surrogates but only the gradient of the penalty). Specifically, SGA like methods can be written as

$$f^{k+1} = f^k + \alpha_k h_k(f^k, \xi_k),$$

where ξ_k is the random subset index, and $h_k(f^k, \xi_k)$ is the search direction, i.e., a preconditioned form of $\nabla\mathcal{L}(f^k)$. OSEM in the additive formulation (4.1) with the search direction $h_k(f^k, t_k)$ is then given by

$$h_k(f^k, t_k) = (f^k \oslash A_{t_k}^\top 1) \odot \nabla\mathcal{L}_{t_k}(f^k).$$

This can viewed as a diagonally preconditioned gradient ascent with respect to the subobjective $\mathcal{L}_{t_k}$. These discussions naturally motivate the following algorithmic developments for the PET MAP problem (3.1). For a given subset index t, we denote

$$\Phi_t(f) := \mathcal{L}_t(f) - \beta N_s^{-1} \mathcal{R}(f).$$

Then the extension of the additive formulation to problem (3.1) leads to SGA updates

$$f_{\text{sga}}^{k+1} = f_{\text{sga}}^k + \alpha_k d_t(f_{\text{sga}}^k) \odot (\nabla\mathcal{L}_{t_k}(f_{\text{sga}}^k) - \beta N_s^{-1} \nabla\mathcal{R}(f_{\text{sga}}^k)), \tag{4.2}$$

where the index t_k is selected uniformly over $[N_s]$, $\alpha_k = 1$, and $d_t(f)$ is given by

$$d_t(f) := f \oslash A_t^\top 1.$$

An update of this type is a standard extension of OSEM algorithms to the MAP problem. However, it is not necessarily maximising the given objective at each step. Note that by including a penalty, the nonnegativity of the updates is generally not preserved, and a projection step by $P_{\geqslant 0}$ applied at each iteration. According to the theory for SGA, the iteration (4.2) generally does not converge to the MAP solution, unless a decaying step-size schedule is employed [13]. This is attributed to the stochasticity of the gradient estimate, and the variance of the estimated ascent direction can significantly slow down the convergence of the algorithm when the iterates approach the maximiser.

One idea to reduce the variance of the gradient estimate in SGA is variance reduction. For the constrained MAP problem, we use proximal versions of SAGA and SVRG [44] to enforce the iterate feasiblity by the projection operator $P_{\geqslant 0}$. Both SAG and SAGA keep a running table of computed gradients of the subobjectives, and then efficiently estimate the full gradient. By rescaling the full gradient, SAGA employs unbiased estimates, whereas the SAG estimate is biased. SAGA estimates of the full gradient are given by

$$\begin{aligned}
q_{t_k}^{k+1} &= \nabla\Phi_{t_k}(f_{\text{saga}}^k), \text{ and for } t \neq t_k \text{ keep } q_t^{k+1} = q_t^k; \\
f_{\text{saga}}^{k+1} &= f_{\text{saga}}^k + \alpha d_{t_k}(f_{\text{saga}}^k) \odot \left(q_{t_k}^{k+1} - q_{t_k}^k + \frac{1}{N_s} \sum_{t=1}^{N_s} q_t^k \right).
\end{aligned} \tag{4.3}$$

SVRG enjoys the convergence rate performance of SAG and SAGA, but does not require maintaining a running list of gradients. Setting $f^{\text{anc}} = f^0$, and $\tilde{q} = \frac{1}{N_s}\nabla\mathcal{L}(f^{\text{anc}}) - \frac{\beta}{N_s}\nabla\mathcal{R}(f^{\text{anc}})$

the algorithm follows

$$f_{\text{svrg}}^{k+1} = f_{\text{svrg}}^k + \alpha d_{t_k}(f_{\text{svrg}}^k) \odot (\nabla\Phi_{t_k}(f_{\text{svrg}}^k) - \nabla\Phi_{t_k}(f^{\text{anc}}) + \tilde{q}).$$

If $k \mod \eta N_s = 0$ set the anchor estimate $f^{\text{anc}} = f_{\text{svrg}}^k$ and update   (4.4)

$$\tilde{q} = N_s^{-1}\nabla\Phi(f^{\text{anc}}).$$

A value of the full gradient update frequency η between 2 and 5 is recommended [26].

Several remarks are in order. First, note that the methods described in sections 2 and 3 aim at explicitly computing the maximiser at each step. However, update equations (4.3) and (4.4) are derived by analogy with the additive formulation (4.1), and thus mathematically less principled. Second, provided that the subobjectives Φ_t are L-Lipschitz, SAG and SAGA converge (sub-linearly in expectation) to the minimiser for the fixed stepsize $\alpha = (16L)^{-1}$ [35]. This result does not apply to the PET problem (3.1) since the subobjectives $\mathcal{L}_t$ are not Lipschitz in a neighbourhood of 0, and since the algorithms use iteration-dependent preconditioners. Third, as observed in [1], preconditioned gradient ascent based algorithms (also known as diagonally-scaled incremental gradient methods) do not always converge to the MAP solution when using iteration-dependent preconditioners. Indeed, assuming $f^k \to f^\star$, that $\nabla\Phi(f)$ is continuous, and that preconditioner functions $d_t(f) \geqslant 0$ are continuous such that $\|d_t(f)\| \neq 0$ for $f \neq 0$, the issue seems to persist for the proposed stochastic algorithms. The latter assumption is satisfied by the EM preconditioner. Assuming $\lim_{k\to\infty}\alpha_k = 0$ and $\sum_{k=1}^{\infty}\alpha_k = \infty$, then the convergence of the SGA algorithm implies

$$\sum_{t=1}^{N_s} d_t(f^\star) \odot \nabla\Phi_t(f^\star) = 0.$$

Thus, unless all the preconditioners d_t are the same, this identity generally differs from the true optimality condition $\nabla\Phi(f^\star) = 0$ (for unconstrained optimisation). Note that an analogous analysis holds for stochastic estimators in expectation, provided the corresponding search direction $h_k(f^k, \xi_k)$ is unbiased and consistent. One simple remedy is to freeze the preconditioner $d_t(f^k) = d$, with $d = d(f^0)$. In practice, gradient-based variance reduction methods, with either iteration dependent or constant preconditioners, perform well for the PET problem [41, 42].

The convergence of non-preconditioned SAGA and SVRG has mostly been studied for strongly convex objectives. Namely, SAGA enjoys linear convergence for strongly convex problems [15], and $O(1/k)$ convergence of the average iterate for convex problems [14, theorem 4.8], whereas the SVRG anchor point converges linearly for strongly convex problems [26]. These assumptions are not satisfied by the PET problem, neither for standard nor modified likelihood (2.11), which is only strictly convex. Nonetheless, combining the arguments in [1, 33] allows establishing almost sure convergence of both SAGA and SVRG for a modified likelihood term and a constant preconditioner. In practice, this holds since the background events are nonzero, for which there is no need to modify the likelihood.

Theorem 4.1. *Let $\tilde{\Phi}(f) = \tilde{\mathcal{L}}(f) - \beta\mathcal{R}(f)$ where $\tilde{\mathcal{L}}(f) = \sum_{i=1}^{M}\varphi_i(a_i^\top f)$ uses the modified likelihood (2.11). Moreover, let all entries of $\mathrm{d} \in \mathbb{R}^N$ be positive, denote by $L = \max_{t\in N_s}L_t$ the largest of the Lipschitz constant of sub-objective gradients $\tilde{\Phi}_t(f)$ and by $d_{\max} = \|d\|_\infty$ the largest entry of d, and assume $\mathrm{argmax}_{f\geqslant 0}\Phi(f) \neq \emptyset$.*

Then taking $\alpha = \frac{1}{3Ld_{\max}^{1/2}}$ and $d_t(f_{\text{saga}}^k) = d$ in the SAGA algorithm (4.3) *we have $\tilde{\Phi}(f_{\text{saga}}^k) \to$ $\Phi(f^\star)$ and $f_{\text{saga}}^k \to f^\star$ almost surely. Taking $\alpha \leqslant \frac{1}{4Ld_{\max}^{1/2}(\eta N_s + 2)}$ and $d_t(f_{\text{svrg}}^k) = d$ in the* svrg *algorithm* (4.4) *we have $f_{\text{svrg}}^k \to f^\star$ almost surely and $\mathbb{E}[\Phi(f^\star) - \tilde{\Phi}(f_{\text{svrg}}^{k\eta N_s})] = \mathcal{O}(1/k)$.*

Note that theorem 4.1 provides conditions for convergence but does not quantify the speed of convergence. This is typical for non-strongly concave problems for most variance reduction techniques. Note that for strongly-concave problems, the convergence rate depends on the ratio between the Lipschitz constant and the stepsize. In case of preconditioned SVRG and SAGA, the effective stepsize is determined not only by α but also by the preconditioner d, as indicated by the bounds on the stepsize in theorem 4.1.

5. Numerical experiments and discussions

In this section, we present numerical results for the two classes of stochastic methods for the MAP problem (3.1) with the log cosh penalty, which is often employed for PET reconstruction. We present two examples: a brain phantom, and a torso XCAT. We examine the performance of SVREM with SVRG and SAGA, both with a constant preconditioner $d = d(f^0)$, which provides consistently better performance than the iteration-dependent counterpart $d(f^k)$. The results for SAG and SARAH were nearly identical with those for SAGA and SVRG, respectively, and thus are not included.
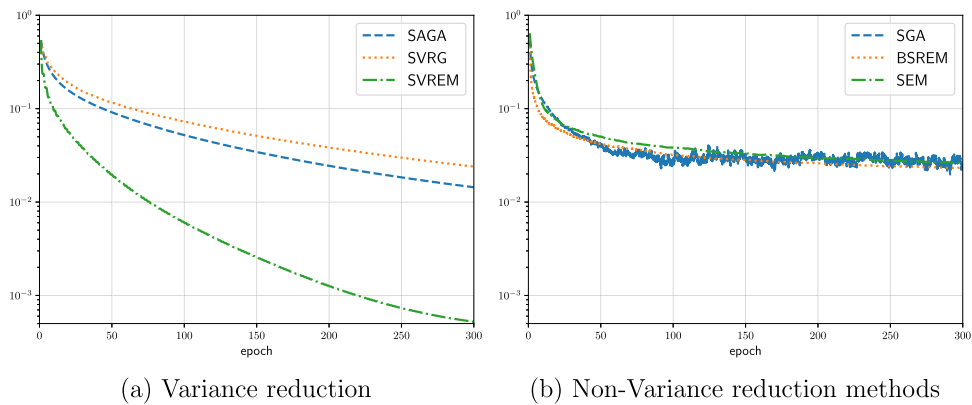
5.1. Brain phantom

In this experiment, we take one slice (of size 114×114) from a BrainWeb phantom [10], available at https://bic.mni.mcgill.ca/brainweb/. The forward map is taken to be the Radon transform using 180 projection angles with a 1 angle separation. The sinogram data is binned into subsets using geometric projections of the scanner. We consider $N_s = 15, 30$, and 45 subsets and accordingly, each subset consists of 12, 6, and 4 views, respectively. In the reconstruction, we use the log cosh penalty with $\delta = 0.01$ and a regularisation parameter $\beta = 60$, which are determined in a trial-and-error manner, and conduct the experiments in MATLAB R2019b. The accuracy of a reconstruction f is measured by the relative error $\Delta(f) = \|f - f^\star\|/\|f^\star\|$, where $f^\star$ is the reference solution, computed using LBFGS-B (available at https://mathworks.com/matlabcentral/fileexchange/35104-lbfgsb-l-bfgs-b-mex-wrapper, retrieved on April 11, 2021). All algorithms are initialised with one epoch of OSEM. Following [41], for SAGA and SVRG, we employ a constant stepsize $\alpha = 2$, and update SVRG with the full gradient every $\eta = 2$ epochs, whereas for SVREM, we choose $\alpha = 0.7$ and $\eta = 1$. We consider three OSEM type methods without variance reduction, i.e., SGA, BSREM, and SEM, which are representative within PET reconstruction, as baselines. More precisely, by SGA we refer to the preconditioned method given in (4.2), which for the full objective Φ reads

$$f_{\text{sga}}^{k+1} = f_{\text{sga}}^k + \alpha(f_{\text{sga}}^k \oslash A_{t_k}^\top 1) \odot \nabla \Phi_{t_k}(f_{\text{sga}}^k).$$

BSREM [1] is an iterative scheme based on OSEM using a decaying stepsize schedule

$$f_{\text{bsrem}}^{k+1} = f_{\text{bsrem}}^k + \alpha_k(f_{\text{bsrem}}^k \oslash A^\top 1) \odot \nabla \Phi_{t_k}(f_{\text{bsrem}}^k).$$

Note that standard BSREM uses a subset-independent preconditioner. We use its subset-dependent variant, replacing $A^\top 1$ with $A_{t_k}^\top 1$ since it exhibits a faster, yet steady, convergence for the example. In the experiments, both SEM (cf (2.9)) and BSREM use the stepsize schedule $\alpha_k = (0.001k + 1)^{-1}$, which is sufficient to ensure their convergence. Since SVRG and

(a) Variance reduction

(b) Non-Variance reduction methods

Figure 1. The error decay behaviour for the brain phantom: (a) SAGA, SVRG, and SVREM, and (b) SGA, BSREM, and SEM.

(a) Reference   (b) SAGA   (c) SVRG   (d) SVREM
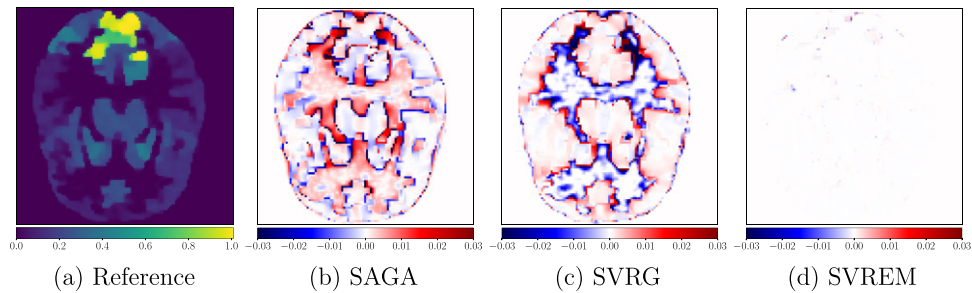
Figure 2. The pixel wise errors of SAGA, SVRG, and SVREM for the brain phantom with respect to the reference solution computed by LBFGS-B.

SVREM require computing the full gradient once every η epochs, to make a fair comparison of overall computational cost, we count epochs in terms of the number of subsets that are used at each iteration. Thus, every ηN_s updates of SVRG and SVREM are counted as $\eta + 1$ epochs, and η epochs of SAGA.

In figure 1, we show the comparative results for all methods on $N_s = 30$ subsets of the data. It is clearly observed that SVREM exhibits the fastest convergence among all the methods, which is also corroborated by the pixel-wise errors in figure 2: the SVREM reconstruction agrees nearly perfectly with the reference solution, whereas the pixelwise errors of SAGA and SVRG reconstructions still clearly exhibit structures, especially edges. Moreover, SAGA is slightly slower than SVRG (in terms of the running time), and both algorithms would benefit from a larger stepsize, though setting it too large impairs the overall convergence. We will explore the stepsize issue for SAGA and SVRG in section 5.2. Just as expected, all variance reduction methods outperform SGA, BSREM, and SEM, and are orders of magnitude faster, especially for high-accuracy solutions, showing clearly the benefit of variance reduction for accelerating OSEM type algorithms. Although not presented, one can observe similar behavior for other subset numbers.

(a) Stepsizes   (b) Update frequency   (c) Number of subsets
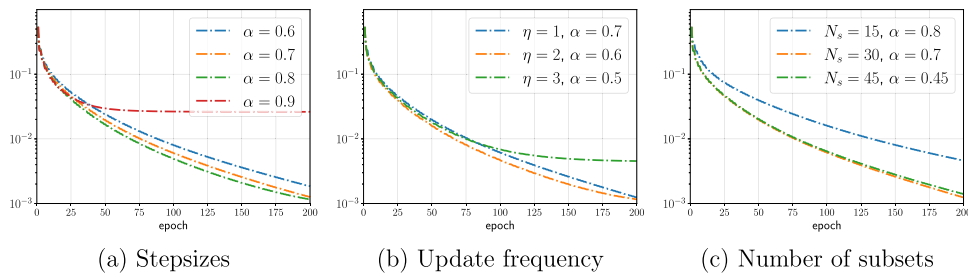
Figure 3. The error decay behaviour for SVREM on the brain phantom with respect to the stepsize α, the update frequency η the number N_s of subsets. In panels (a) and (b) we use 30 subsets of the data.

Table 2. Comparison of average running times (in seconds) for experiments in section 5.1.

Method\N_s	15	30	45
SGA	15.31	25.80	36.64
BSREM	15.28	25.86	36.77
SEM	16.15	27.26	38.42
SAGA	16.66	28.25	40.14
SVRG	17.20	29.06	40.49
SVREM	20.13	33.34	47.01

Figure 3 studies the convergence behavior of SVREM with respect to three important algorithmic parameters, i.e., update frequency η (of full gradient updates), the stepsize α, and number of subsets N_s. Generally, all these parameters greatly impact the performance of SVREM, and they should be tuned together to achieve optimal convergence behavior. It is observed that increasing the frequency η provides some acceleration in initial epochs but a too large η can impair the asymptotic convergence, even with a tuned stepsize α. The stepsize α has a similar influence as the frequency η: a larger stepsize α gives faster initial acceleration, but too large a value may prevent the algorithm from converging to the MAP maximiser. Lastly, a larger number N_s of subsets (with suitably tuned stepsizes) tends to provide faster initial convergence. These observations show clearly the importance of proper partition of the subsets.

In table 2 we show the runtime, averaged over 10 independent runs, for running 10 epochs of the algorithms. Note that the implementations have not been optimised, e.g., computation of the coefficients in (3.4), and thus the numbers should be interpreted indicatively only. It is observed that all the methods have comparable runtime, and that methods that require storing an estimator of the full gradient are a bit slower than the methods for which this is not required, e.g., SEM versus SVREM, or BSREM versus SVRG. This observation holds when the number N_s of subsets is fixed. In theory the runtime per epoch should be independent of the number of subsets ignoring the computation of the penalty; the results in table 2 actually show an increase with increasing N_s which we attribute to the extra overhead in multiple function calls as opposed to an ideally vectorised implementation.
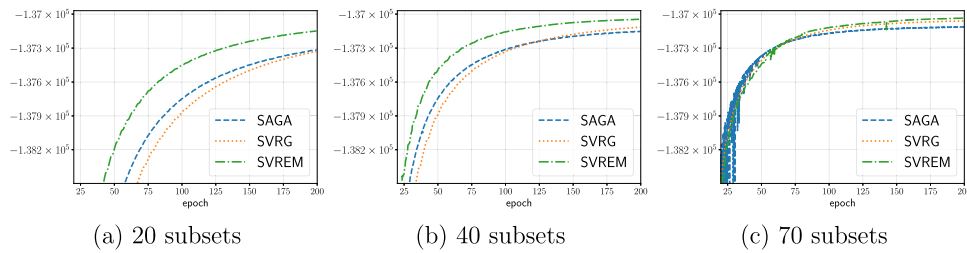
Figure 4. Performance comparison of SAGA, SVRG, and SVREM for $20, 40$, and 70 subsets over 200 epochs.
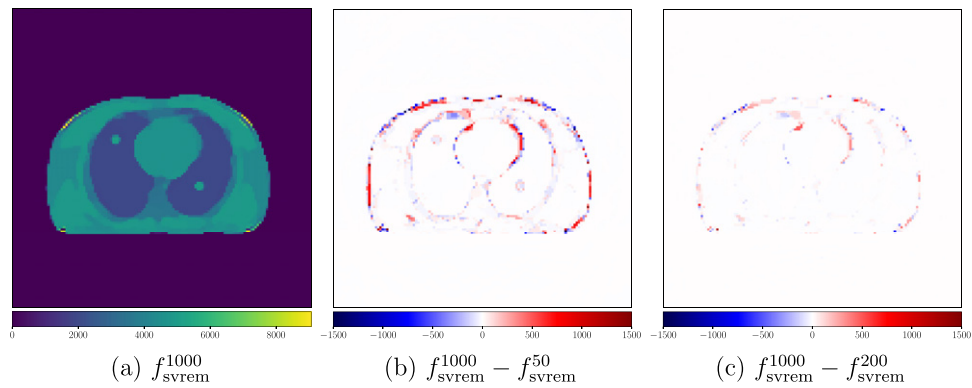
5.2. Torso phantom

To evaluate the algorithms in a more realistic PET setting, in this experiment we use a PET scan of a torso, obtained as an XCAT simulated phantom [36] with 2 rings and 280 projection angles. The sinogram data consists of 280 views, and is binned into 20, 40, and 70 subsets using geometric projections of the scanner, with each subset having 14, 7, and 4 views, respectively. In the experiment, SVRG and SVREM update the full gradient and full expected statistic, once every $\eta = 5$ and $\eta = 3$ epochs, respectively, and all algorithms are initialised with one epoch of OSEM. The reconstruction was carried out using the Software for Tomographic Image Reconstruction [39], via a python environment, available at https://github.com/UCL/STIR. We use the log cosh penalty with $\delta = 1$ and fix $\beta = 0.0001$. Since a reference solution is unavailable, we evaluate the accuracy by the objective value.

The results for 200 epochs of SAGA, SVRG and SVREM are shown in figure 4. Unlike the brain phantom, the behaviour with respect to η is more stable, and choosing a larger η provides a better per-epoch comparison with SAGA. One interesting observation is about the feasible stepsize regime for SVREM. The work [9] suggests $\alpha \in (0, 1)$, with an upper bound depending on the Lipschitz constant of subset statistics (and also the number of subsets). This behaviour was also observed in figure 3(a). In contrast, for the torso phantom, the admissible stepsize seem to be $(0, 2]$, which corresponds to over-relaxation, a well-known practice in iterative linear solvers [43] (see [40] for an application in EM algorithm). In figure 4, we use $\alpha = 2.0$ for 20 and 40 subsets, and $\alpha = 1.4$ for 70 subsets. In this challenging setting, SAGA, SVRG and SVREM show comparable overall performance, cf figure 4. Nonetheless, SVREM consistently outperforms SVRG and SAGA, especially when the number N_s of subsets is small. Moreover, SAGA occasionally exhibits an undesirable stochastic behaviour during initial iterations (most notably for 70 subsets), before stabilising after roughly 70 epochs. The latter can be remedied by choosing a smaller stepsize α, which of course can adversely affect the overall convergence behavior. These results suggest that SVREM provides a greater benefit for a smaller number N_s of subsets, and as the number N_s of subsets increases the algorithms become comparable.

In figure 5 we study the convergence behaviour over a longer epoch horizon, using 40 subsets and 1000 epochs, and examine how do SAGA, SVRG, and SVREM depend on the stepsize α. The numerical results show that SVREM with $\alpha = 1.0$ (and also for smaller values) can eventually outperform both SVRG and SAGA, and catches up to SVREM with a larger stepsize, confirming the asymptotic convergence. Meanwhile the results for SVRG, and particularly SAGA, suggest that a smaller stepsize α is needed to ensure their convergence, since otherwise the iterations enter a limit cycle. Thus, reducing the stepsize has a conflicting effect: it slows down the speed of convergence during initial iterations, but it provides better asymptotic behaviour. For a more effective profile of the overall convergence, a dynamically

(a) SAGA and SVRG   (b) SVREM
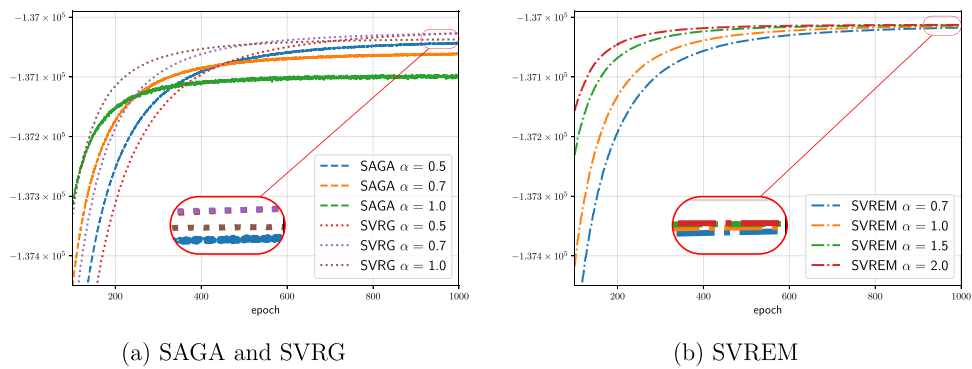
Figure 5. The influence of the stepsize α on SAGA, SVRG, and SVREM for the XCAT torso with 40 subsets and over 1000 epochs.

(a) $f_{\mathrm{svrem}}^{1000}$ (b) $f_{\mathrm{svrem}}^{1000} - f_{\mathrm{svrem}}^{50}$ (c) $f_{\mathrm{svrem}}^{1000} - f_{\mathrm{svrem}}^{200}$

Figure 6. (a) SVREM reconstruction after 1000, and (b) and (c) pixel-wise differences of SVREM reconstructions after 200 and 50 epochs.

variable stepsize schedule emerges as a natural choice. To gain further insights, in figure 6 we show pixel-wise differences for SVREM with 40 subsets. The results show that the background and the smooth parts of the image are mostly resolved after 50 epochs, whereas the edges are more challenging to resolve and are still improving after 50 epochs.

6. Conclusions

In this work we revisited two classes of stochastic EM algorithms based on variance reduction from the machine learning community for the penalised PET reconstruction, one from the perspective of EM algorithms and the other from the perspective of preconditioned SGA. Both classes of algorithms are straightforward to implement, and are applicable to a wide range of penalty terms. The numerical results indicate that these algorithms can effectively and efficiently accelerate the convergence of classical OSEM type algorithms, and hold significant potentials for the MAP problem in PET reconstructions. This is particularly true for SVREM, which enjoys steady convergence towards the maximising solution, sometimes even

with over-relaxation. These promising empirical results naturally motivate further experimental evaluations on patient data and a theoretical analysis, to investigate the interplay between the number of subsets and the stepsize schedule (and also the 'full-gradient' update frequency for SVRG), the role of over-relaxation in SVREM, and to precisely characterise their influence on the convergence speed.

Acknowledgments

Data availability statement

No new data were created or analysed in this study.

Appendix A. Proof of theorem 4.1

Note that the constrained maximisation problem

$$\underset{f \geqslant 0}{\mathrm{argmax}} \Phi(f), \quad \text{with} \quad \Phi(f) := \mathcal{L}(f) - \beta \mathcal{R}(f),$$

can be written as a minimisation problem

$$\arg \min_f - \Phi(f) + \chi_{\geqslant 0}(f),$$

where $\chi_{\geqslant 0}(f)$ is the indicator function of vectors with nonnegative entries. It is proper, convex and lower semi-continuous and its proximal operator is the projection $P_{\geqslant 0}$. The conditions of the theorem imply

(a) The preconditioners $d(f^k) = d$ are constant, diagonal and positive.

(b) $\Phi : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable with a Lipschitz gradient, and Φ_t are continuously differentiable with an L_t-Lipschitz gradient for each $t \in [N_s]$. This may require bounding the projection operator to a finite box, whose bounds can be explicitly computed [1] and which in practice affects neither the convergence nor the iterations.

(c) The set of maximisers is nonempty, i.e., $\mathrm{argmax}_{f \geqslant 0} \Phi(f) \neq \emptyset$.

Then assumption (b) holds for the modified objective $\tilde{\Phi}(f) := \mathcal{L}(f) - \beta \mathcal{R}(f)$, and the parameter $\varepsilon > 0$ in (2.11) can be chosen to be sufficiently small so that [1] $\mathrm{argmax}_{f \geqslant 0} \Phi(f) = \mathrm{argmax}_{f \geqslant 0} \tilde{\Phi}(f)$ and $\max_{f \geqslant 0} \Phi(f) = \max_{f \geqslant 0} \tilde{\Phi}(f)$. Below we use $\tilde{\Phi}$ and identify $\tilde{\Phi} = \Phi$. Motivated by [1, appendix C], we rewrite SAGA as

$$\tilde{f}^k_{\mathrm{saga}} = P_{\geqslant 0} \left(\tilde{f}^{k-1}_{\mathrm{saga}} + \alpha \left(\tilde{q}^k_{t_k} - \tilde{q}^{k-1}_{t_k} + \frac{1}{N_s} \sum_{t=1}^{N_s} \tilde{q}^{k-1}_t \right) \right),$$

with

$$\tilde{f}^k = d^{-1/2} \odot f^k, \qquad \tilde{q}^k_t = \nabla \Psi_t(\tilde{f}^k), \quad \text{for} \quad \Psi_t(\tilde{f}) = \Phi_t(d^{1/2} \odot \tilde{f}).$$

Now let $L = \max_{t \in [N_s]} L_t$. Since Φ_t are L-Lipschitz and d has non-negative entries, Ψ_t are also Lipschitz, with a Lipschitz constant $L\|d^{1/2}\|_\infty = Ld_{\max}^{1/2}$. Then with $\alpha = (3Ld_{\max}^{1/2})^{-1}$, by [33, theorem 2.1] (by changing maximisation to minimisation), there exists an $\tilde{f}^\star \in \arg\max_{\tilde{f} \geqslant 0} \Psi(\tilde{f})$ such that $\Psi(\tilde{f}_{\text{saga}}^k) \to \Psi(\tilde{f}^\star)$ and $\tilde{f}_{\text{saga}}^k \to \tilde{f}^\star$ almost surely.

Similarly for SVRG, we have

$$\tilde{f}_{\text{svrg}}^k = P_{\geqslant 0}(\tilde{f}_{\text{svrg}}^{k-1} + \alpha(\nabla\Psi_{t_k}(\tilde{f}_{\text{svrg}}^{k-1}) - \nabla\Psi_{t_k}(\tilde{f}^{\text{anc}}) + \tilde{G})),$$

where $\tilde{G} = \frac{1}{N_s}\nabla\Psi(\tilde{f}^{\text{anc}})$. Then by [33, theorem 2.2.(i)], with $\alpha \leqslant (4Ld_{\max}^{1/2}(\eta N_s + 2))^{-1}$, there exists an $\tilde{f}^\star \in \arg\max_{\tilde{f} \geqslant 0} \Psi(\tilde{f})$ such that $\tilde{f}_{\text{svrg}}^k \to \tilde{f}^\star$ almost surely. Moreover, at the point when the full gradient is updated, i.e. if $k \bmod \eta N_s = 0$, we have $\mathbb{E}[\Psi(\tilde{f}^\star) - \Psi(\tilde{f}_{\text{svrg}}^k)] = \mathcal{O}(k^{-1})$. By definition, $\Psi(\tilde{f}) = \Phi(f)$, we have $\Phi(f_{\text{saga}}^k) \to \Phi(f^\star)$ and $f_{\text{saga}}^k \to f^\star$, and $f_{\text{svrg}}^k \to f^\star$ almost surely. Moreover, we have $\mathbb{E}[\Phi(f^\star) - \Phi(f_{\text{svrg}}^{(k\eta N_s)})] = \mathcal{O}(k^{-1})$.

Appendix B. The differences between SVREM and SVRG for ML problems

It is tempting to think that SVRG and SVREM are equivalent in a certain context. We compare the two algorithms for the simplest case, i.e., the ML problem. Assume that both algorithms employ the same anchored estimate f^{anc}, the full gradient, and the subset index t. For SVREM, we first update the estimator $\widehat{s}^k$ of the full statistic by

$$\widehat{s}^{k+1} = (1 - \alpha)\widehat{s}^k + \alpha\left(\tau_{t_k}(\widehat{f}_{\text{svrem}}^k) - \tau_{t_k}(\widehat{f}^{\text{anc}}) + \widehat{s}(f^{\text{anc}})\right),$$

and then update the maximising solution f_{svrem}^k by

$$f_{\text{svrem}}^{k+1} = \left(\frac{\widehat{s}^{k+1}}{\sum_{m=1}^M a_{mn}}\right)_{n=1}^N = f_{\text{svrem}}^k + \alpha h^k,$$

where the search direction h^{k+1} is given by

$$N_s^{-1} h^{k+1} = d(f^k) \odot \nabla\mathcal{L}_t(f^k) - d(f^{\text{anc}}) \odot \nabla\mathcal{L}_t(f^{\text{anc}})$$
$$+ N_s^{-1} d(f^{\text{anc}}) \odot \nabla\mathcal{L}(f^{\text{anc}})$$
$$+ (f^k - f^{\text{anc}}) \odot (A_t^\top 1 \oslash A^\top 1 - N_s^{-1}).$$

Meanwhile, for SVRG updates with variable preconditioner, we have

$$f_{\text{svrg}}^{k+1} = f_{\text{svrg}}^k + \alpha d(f_{\text{svrg}}^k) \odot (\nabla\mathcal{L}_{t_k}(f_{\text{svrg}}^k) - \nabla\mathcal{L}_{t_k}(f^{\text{anc}}) + N_s^{-1}\nabla\mathcal{L}(f^{\text{anc}})).$$

Thus, the two algorithms differ greatly even if SVRG uses a iteration dependent preconditioner $d(f_{\text{svrg}}^k) = f_{\text{svrg}}^k \oslash A^\top 1$ (or a subset dependent preconditioner). There are two notable differences. First, SVREM uses different preconditioners for each term, i.e., the gradient terms containing f^{anc} are preconditioned with $d(f^{\text{anc}})$, and $d(f^k)$, for SVREM and SVRG, respectively. Second, the search direction for SVREM involves a term akin to momentum.

ORCID iDs

Željko Kereta ⓘ https://orcid.org/0000-0003-2805-0037
Simon Arridge ⓘ https://orcid.org/0000-0003-1292-0210
Kris Thielemans ⓘ https://orcid.org/0000-0002-5514-199X
Bangti Jin ⓘ https://orcid.org/0000-0002-3775-9155

References

[1] Ahn S and Fessler J A 2003 Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms *IEEE Trans. Med. Imaging* **22** 613–26
[2] Bredies K and Holler M 2020 Higher-order total variation approaches and generalisations *Inverse Problems* **36** 123001
[3] Browne J and de Pierro A B 1996 A row-action alternative to the EM algorithm for maximizing likelihood in emission tomography *IEEE Trans. Med. Imaging* **15** 687–99
[4] Byrne C L 1998 Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods *IEEE Trans. Image Process.* **7** 100–9
[5] Cappé O and Moulines E 2009 On-line expectation-maximization algorithm for latent data models *J. R. Stat. Soc.* B **71** 593–613
[6] Chambolle A, Ehrhardt M J, Richtárik P and Schönlieb C-B 2018 Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications *SIAM J. Optim.* **28** 2783–808
[7] Chambolle A and Pock T 2011 A first-order primal-dual algorithm for convex problems with applications to imaging *J. Math. Imaging Vis.* **40** 120–45
[8] Chang J-H, Anderson J M M and Votaw J R 2004 Regularized image reconstruction algorithms for positron emission tomography *IEEE Trans. Med. Imaging* **23** 1165–75
[9] Chen J, Zhu J, Teh Y W and Zhang T 2018 Stochastic expectation maximization with variance reduction *NeurIPS* pp 7967–77
[10] Cocosco C A, Kollokian V, Kwan R K-S, Pike G B and BrainWeb A C E 1997 Online interface to a 3D MRI simulated brain database *NeuroImage* **5** 425
[11] de Pierro A R 1993 On the relation between the ISRA and the EM algorithm for positron emission tomography *IEEE Trans. Med. Imaging* **12** 328–33
[12] de Pierro A R 1995 A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography *IEEE Trans. Med. Imaging* **14** 132–7
[13] de Pierro A R and Yamagishi M E B 2001 Fast EM-like methods for maximum *a posteriori* estimates in emission tomography *IEEE Trans. Med. Imaging* **20** 280–8
[14] Defazio A 2014 New optimization methods for machine learning *PhD Thesis* Australian National University
[15] Defazio A, Bach F and Lacoste-Julien S 2014 SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives *NeurIPS* pp 1646–54
[16] Dempster A P, Laird N M and Rubin D B 1977 Maximum likelihood from incomplete data via the EM algorithm *J. R. Stat. Soc.* B **39** 1–22
[17] Ehrhardt M J, Markiewicz P and Schönlieb C-B 2019 Faster PET reconstruction with non-smooth priors by randomization and preconditioning *Phys. Med. Biol.* **64** 225019
[18] Erdogan H and Fessler J A 1998 A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction *IEEE Nuclear Science Symp. Medical Imaging Conf. (NSS/MIC)*
[19] Fessler J A and Hero A O 1995 Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms *IEEE Trans. Image Process.* **4** 1417–29
[20] Gower R M, Schmidt M, Bach F and Richtarik P 2020 Variance-reduced methods for machine learning *Proc. IEEE* **108** 1968–83
[21] Green P J 1990 Bayesian reconstructions from emission tomography data using a modified EM algorithm *IEEE Trans. Med. Imaging* **9** 84–93
[22] Herman G T and Meyer L B 1993 Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application) *IEEE Trans. Med. Imaging* **12** 600–9
[23] Hohage T and Werner F 2016 Inverse problems with Poisson data: statistical regularization theory, applications and algorithms *Inverse Problems* **32** 093001

[24] Hudson H M and Larkin R S 1994 Accelerated image reconstruction using ordered subsets of projection data *IEEE Trans. Med. Imaging* **13** 601–9

[25] Ito K and Jin B 2015 *Inverse Problems: Tikhonov Theory and Algorithms* (Singapore: World Scientific)

[26] Johnson R and Zhang T 2013 Accelerating stochastic gradient descent using predictive variance reduction *NeurIPS* pp 314–23

[27] Karimi B, Wai H-T, Moulines E and Lavielle M 2019 On the global convergence of (fast) incremental expectation maximization methods *NeurIPS* pp 2837–47

[28] Kaufman L 1987 Implementing and accelerating the EM algorithm for positron emission tomography *IEEE Trans. Med. Imaging* **6** 37–51

[29] Lange K and Carson R E 1984 EM reconstruction algorithms for emission and transmission tomography *J. Comput. Tomo.* **8** 306–16

[30] Lange K, Hunter D R and Yang I 2000 Optimization transfer using surrogate objective functions *J. Comput. Graphic. Stat.* **9** 1–59 With discussion, and a rejoinder by Hunter and Lange

[31] Neal R M and Hinton G E 1999 A view of the EM algorithm that justifies incremental, sparse, and other variants *Learning in Graphical Models* ed M I Jordan (Cambridge, MA: MIT Press) pp 355–68

[32] Nuyts J, Bequé D, Dupont P and Mortelmans L 2002 A concave prior penalizing relative differences for maximum-*a posteriori* reconstruction in emission tomography *IEEE Trans. Nucl. Sci.* **49** 56–60

[33] Poon C, Liang J and Schönlieb C-B 2018 Local convergence properties of SAGA/prox-SVRG and acceleration *ICML* pp 4124–32

[34] Qi J and Leahy R M 2006 Iterative reconstruction techniques in emission computed tomography *Phys. Med. Biol.* **51** R541–78

[35] Schmidt M, Le Roux N and Bach F 2017 Minimizing finite sums with the stochastic average gradient *Math. Program.* **162** 83–112

[36] Segars W P, Sturgeon G, Mendonca S, Grimes J and Tsui B M W 2010 4D XCAT phantom for multimodality imaging research *Med. Phys.* **37** 4902–15

[37] Shepp L A and Vardi Y 1982 Maximum likelihood reconstruction for emission tomography *IEEE Trans. Med. Imaging* **1** 113–22

[38] Shepp L A, Vardi Y and Kaufman L 1985 A statistical model for positron emission tomography *J. Am. Stat. Assoc.* **80** 8–37

[39] Thielemans K, Tsoumpas C, Mustafovic S, Beisel T, Aguiar P, Dikaios N and Jacobson M W 2012 STIR: software for tomographic image reconstruction release 2 *Phys. Med. Biol.* **57** 867–83

[40] Tsai Y-J, Huang H-M, Fang Y-H D, Chang S-I and Hsiao I-T 2015 Acceleration of MAP-EM algorithm via over-relaxation *Comput. Med. Imaging Graph.* **40** 100–7

[41] Twyman R, Arridge S, Jin B, Hutton B F, Brusaferri L and Thielemans K 2020 Stochastic variance reduction optimisation algorithms applied to iterative PET reconstruction *2020 IEEE Nuclear Science Symp. Medical Imaging Conf. (NSS/MIC)* pp 1–2

[42] Twyman R, Arridge S and Thielemans K 2021 Investigation of subset methodologies applied to penalised iterative PET reconstruction *2021 Int. Meeting on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine* (unpublished)

[43] Varga R S 2000 *Matrix Iterative Analysis* expanded edn (Berlin: Springer)

[44] Xiao L and Zhang T 2014 A proximal stochastic gradient method with progressive variance reduction *SIAM J. Optim.* **24** 2057–75

[45] Zhu R, Wang L, Zhai C and Gu Q 2017 High-dimensional variance-reduced stochastic gradient expectation-maximization algorithm *ICML* pp 4180–8