Fast PET Reconstruction with Variance Reduction and Prior-Aware Preconditioning

Matthias J. Ehrhardt [1], Zeljko Kereta [2], and Georg Schramm [3]

[1]Department of Mathematical Sciences, University of Bath, UK
[2]Computer Science Department, University College London, UK
[3]Department of Imaging and Pathology, KU Leuven, Belgium

June 12, 2025

Abstract: We investigate subset-based optimization methods for positron emission tomography (PET) image reconstruction incorporating a regularizing prior. PET reconstruction methods that use a prior, such as the relative difference prior (RDP), are of particular relevance, as they are widely used in clinical practice and have been shown to outperform conventional early-stopped and post-smoothed ordered subsets expectation maximization (OSEM).

Our study evaluates these methods on both simulated data and real brain PET scans from the 2024 PET Rapid Image Reconstruction Challenge (PETRIC), where the main objective was to achieve RDP-regularized reconstructions as fast as possible, making it an ideal benchmark. Our key finding is that incorporating the effect of the prior into the preconditioner is crucial for ensuring fast and stable convergence.

In extensive simulation experiments, we compare several stochastic algorithms—including Stochastic Gradient Descent (SGD), Stochastic Averaged Gradient Amelioré (SAGA), and Stochastic Variance Reduced Gradient (SVRG)—under various algorithmic design choices and evaluate their performance for varying count levels and regularization strengths. The results show that SVRG and SAGA outperformed SGD, with SVRG demonstrating a slight overall advantage. The insights gained from these simulations directly contributed to the design of our submitted algorithms, which formed the basis of the winning contribution to the PETRIC 2024 challenge.

1 Introduction

1.1 Context

PET is a pillar in modern clinical imaging widely used in oncology, neurology and cardiology. Most state-of-the-art approaches for the image reconstruction problem in PET imaging can be cast as an optimization problem

$$x^* \in \arg\min_x \{\mathcal{D}(Ax + r, y) + \mathcal{R}(x)\}, \tag{1}$$

where the data-fidelity term $\mathcal{D} : \mathcal{Y} \times \mathcal{Y} \to [0, \infty]$ measures how well the estimated data $Ax + r$ matches the acquired data y and the regularizer $\mathcal{R} : \mathcal{X} \to [0, \infty]$ penalizes unwanted features in the image. $A : \mathcal{X} \to \mathcal{Y}$ is a linear forward model for the PET physics, including effects such as scanner sensitivities or attenuation, and r is the additive background term to account for scattered and random coincidences. Due to the Poisson nature of the data, the data-fidelity is usually taken as the Kullback–Leibler (KL) divergence. The regularizer commonly entails nonnegativity constrains and terms promoting smoothness. A particularly successful model for smoothness in PET is the RDP [1].

This paper is concerned with algorithms for a fast reconstruction of x^*. Particularly, we present our winning contribution to the 2024 image reconstruction challenge PETRIC [2], where the task was to reconstruct data acquired by a range of PET scanners using RDP regularized reconstruction methods. PET image reconstructions that use the RDP are of particular current relevance, as RDP is widely used in clinical practice, being implemented by a major commercial vendor, and has been

shown to outperform conventional early-stopped and post-smoothed OS-MLEM reconstructions [3, 4, 5]. Their implementation is based on BSREM [6]. It was shown to be outperformed in terms of speed by an algorithm using ideas from machine learning and a tailored preconditioning [7]. In this paper, we outline our process in finding the winning algorithm and share the insights we gained along the way. For context, the task had to be completed in Synergistic Image Reconstruction Framework (SIRF) [8] and speed was measured as walltime until an application-focused convergence criteria were reached.

1.2 Problem Details

Fast algorithms for PET reconstruction have traditionally been subset-based [9], that is, only a subset of the data is used in every iteration. Over the last decade, algorithms using a similar strategy but derived for machine learning have entered the field, showing state-of-the-art performance [7, 10, 11, 12, 13, 14]. They exploit the fact that the KL divergence is separable in the estimated data

$$\mathcal{D}(Ax + r, y) = \sum_{i=1}^{n} \sum_{j \in S_i} d(A_j x + r_j, y_j), \tag{2}$$

where n denotes the number of subsets and function d is defined by

$$d(s,t) = \begin{cases} s - t + t\log(t/s), & \text{if } t > 0, s > 0 \\ s, & \text{if } t = 0, s \geq 0 \\ \infty, & \text{otherwise} \end{cases}.$$

Here S_i denote a subset of the data, e.g., all data associated to a "view".

A lot of effort has been put into finding good prior models (i.e., regularizers) for PET, including smooth and nonsmooth priors, promoting smoothness of the image to be reconstructed or promoting similarity to anatomical information [15, 16, 17, 18, 19]. In [1], the authors propose a smooth and convex prior that takes into account the scale of typical PET images, resulting in promoting more smoothness in less active regions. Mathematically for nonnegative images x the resulting regularizer can be defined by

$$\mathcal{S}(x) = \frac{1}{2} \sum_{i} \sum_{j \in N_i} w_{i,j} \kappa_i \kappa_j \frac{(x_i - x_j)^2}{x_i + x_j + \gamma |x_i - x_j| + \varepsilon}, \tag{3}$$

where the first sum is over all voxels i and the second sum is over all "neighbors" j. The parameter $\gamma > 0$ allows placing more or less emphasis on edge-preservation and the parameter $\varepsilon > 0$ ensures that the function is well-defined and twice continuously differentiable. The terms $w_{i,j}$, κ_i and κ_j are weight factors accounting for distances between voxels and are intended to create a uniform "perturbation response" [20]. Note that the essential part of the prior is

$$\phi(s,d) = \frac{d^2}{s + \gamma|d| + \varepsilon},$$

which has two important properties. First, if the sum of activities between voxels s is small compared to the scaled absolute difference $\gamma|d|$, the regularizer essentially reduces to total variation: $\phi(s,d) \approx |d|/\gamma$. Second, the larger the activity in both voxels, i.e., the larger s, the less weight is given on penalizing their difference, justifying the name of the regularizer. See also Appendix A.1 for formulas of derivatives.

Combined with the indicator function of the nonnegativity constraint,

$$\iota_{\geq 0}(x) = \begin{cases} 0, & \text{if } x_i \geq 0 \text{ for all } i \\ \infty, & \text{otherwise} \end{cases},$$

we arrive at the regularization model used in PETRIC

$$\mathcal{R}(x) = \beta \mathcal{S}(x) + \iota_{\geq 0}(x). \tag{4}$$

This formula has to be interpreted to be ∞ for infeasible images with negative voxel-values and has the finite RDP value everywhere else.

The rest of the paper is structured as follows. In section 2 we introduce the building blocks of our algorithms. We discuss proximal stochastic gradient approaches for the solution of (1), the stepsize regimes, preconditioning and subset selection. In section 3 we thoroughly investigate the effects of different choices of building blocks in a simulated setting. In section 4 we present the algorithms we ended up using in PETRIC and their performance on real data. We conclude in sections 5 and 6 with final remarks.

2 Building Blocks

Combining the modeling choices in (1), (2) and (4), we arrive at the optimization problem

$$\min_x \left\{ \sum_{i=1}^{n} \mathcal{J}_i(x) + \iota_{\geq 0}(x) \right\}, \tag{5}$$

where we define $\mathcal{J}_i(x) = \mathcal{D}_i(x) + \frac{\beta}{n}\mathcal{S}(x)$, and $\mathcal{D}_i(x) \coloneqq \sum_{j \in S_i} d(A_j x + r_j, y_j)$. The zoo of optimization methods for solving instances of problem (5) is rich and has been growing in recent decades, see [13] and references therein. For linear inverse problems, such as in PET image reconstruction, the most common approaches are based on (proximal) gradient descent or on primal-dual approaches.

In this work we consider stochastic gradient methods for the solution of the problem (5). They take the form

$$x^{(k+1)} = \operatorname{prox}_{\iota_{\geq 0}}\left(x^{(k)} - \tau^{(k)} D^{(k)} \tilde{\nabla}^{(k)}\right), \tag{6}$$

where $\tau^{(k)} > 0$ is a stepsize, $\tilde{\nabla}^{(k)}$ is an estimator of the gradient of the smooth part of the objective function $\mathcal{J}(x) = \sum_{i=1}^{n} \mathcal{J}_i(x)$, $D^{(k)}$ is a matrix that acts as a preconditioner, and $\operatorname{prox}_{\iota_{\geq 0}}$ is the proximal operator associated with the nonnegativity constraint which can be efficiently computed entry-wise, $[\operatorname{prox}_{\iota_{\geq 0}}(x)]_j = \max(0, x_j)$.

All three components $\tilde{\nabla}^{(k)}$, $D^{(k)}$ and $\tau^{(k)}$ are critical for fast and stable algorithmic performance. In realistic image reconstruction settings, and in the context of the PETRIC challenge, the selection of these three components must balance accuracy and computational costs. In the remainder of this section, we review stochastic estimators and discuss their tradeoffs, address the stepsize selection and preconditioners. Lastly, we consider the role of subset selection and sampling regimes. Namely, how to choose the sets S_i in (2) and decide which subsets to use at each iteration of the algorithm.

2.1 Stochastic Gradient Methods

Let's turn our attention to the selection of gradient estimators $\tilde{\nabla}^{(k)}$.

Stochastic Gradient Descent (SGD) defines the gradient estimator by selecting a random subset index i_k in each iteration and evaluating

$$\tilde{\nabla}^{(k)} := n \nabla \mathcal{J}_{i_k}(x^{(k)})$$

to compute the update (6). Each iteration requires storing only the current iterate and computing the gradient of only one subset function. This can lead to large variances across updates, which increase with the number of subsets. To moderate this, vanishing stepsizes, satisfying

$$\sum_{k=1}^{\infty} \tau^{(k)} = \infty \text{ and } \sum_{k=1}^{\infty} (\tau^{(k)})^2 < \infty,$$

are required to ensure convergence, but at the cost of convergence speed.

Stochastic Averaged Gradient Amelioré (SAGA) [21] controls the variance by keeping a table of historical gradients $(g_i^{(k)})_{i=1}^n \in \mathcal{X}^n$. Each iteration uses a computed subset gradient combined with the full gradient table to update the gradient estimator

$$\tilde{\nabla}^{(k)} = n(\nabla \mathcal{J}_{i_k}(x^{(k)}) - g_{i_k}^{(k)}) + \sum_{i=1}^{n} g_i^{(k)}, \tag{7}$$

followed by updating the corresponding entry in the table

$$g_j^{(k+1)} = \begin{cases} \nabla \mathcal{J}_{i_k}(x^{(k)}), & \text{if } j = i_k \\ g_j^{(k)}, & \text{otherwise} \end{cases}.$$

In contrast to SGD, SAGA guarantees convergence to a minimizer with constant stepsizes and preconditioners for Lipschitz-smooth problems. In its standard form SAGA has the same computational cost as SGD, but requires storing n gradients. The memory cost is not a practical limitation for most PET problems (even for relatively large n). If this is a concern, alternative formulations of SAGA exist with other memory footprints, see [13] for a further discussion.

Stochastic Variance Reduced Gradient (SVRG) [22] reduces the variance by storing reference images and gradients. In contrast to SAGA, these are updated infrequently. SVRG is usually implemented with two loops: an outer loop and an inner loop. At the start of each outer loop subset gradients and the full gradient estimator are computed at the last iterate as

$$\hat{g}_i = \nabla \mathcal{J}_i(\hat{x}), \quad \hat{g} = \sum_{i=1}^{n} \hat{g}_i.$$

In the inner loop the gradients are retrieved from memory and balanced against a randomly sampled subset gradient at the current iterate, giving the gradient estimator

$$\tilde{\nabla}^{(k)} = n(\nabla \mathcal{J}_{i_k}(x^{(k)}) - \hat{g}_{i_k}) + \hat{g}. \tag{8}$$

Note the similarity between the gradient estimators of SAGA and SVRG given by (7) and (8), respectively. After ωn iterations the snapshot image and the full gradient estimator are updated. The update parameter $\omega \in \mathbb{N}$ is typically chosen as 2 for convex problems.

It is most common to store only the snapshot image $\hat{x}$ and the corresponding full gradient $\sum_{i=1}^{n} \hat{g}_i$, which then requires recomputing the subset gradient $\hat{g}_{i_k}$ at each iteration. This lowers the memory footprint (requiring only the snapshot image and the full gradient to be stored), but increases the computational costs.

2.2 Stepsizes

Theoretical convergence guarantees often require stepsizes based on $L_{\max} = \max_{i=1,\dots,n}\{L_i\}$, where L_i is the Lipschitz constant of $\nabla \mathcal{J}_i$. In PET, global Lipschitz constants are usually pessimistic, yielding conservative stepsize estimates.

Many stepsize approaches exist for stochastic iterative methods, ranging from predetermined choices made before running the algorithm (constant or vanishing), to adaptive methods (e.g., Barzilai–Borwein (BB) [23] and "difference of gradients"-type [24] rules), and backtracking techniques (e.g., Armijo [25]). Due to the constraints imposed by the challenge (where computational time is a key metric), in this work we focus on the first two categories.

Constant is the baseline stepsize rule. The specific value requires tuning to ensure convergence.

Vanishing rules consider stepsizes of the form $\tau^{(k)} = \tau^{(0)}/(1 + \eta k/n)$, which satisfy SGD convergence conditions, for $\tau^{(0)} > 0$ and decay parameter $\eta > 0$ that needs balance convergence and stability: small enough to maintain speed but large enough to ensure convergence.

Adaptive stepsize tuning via the BB rule is achieved by minimizing the residual of the secant equation at the current iterate. It converges for strongly convex problems and it is applicable to SGD and SVRG [23]. We experimented with several variants (long and short forms, geometric mean combinations, diagonal BB, etc.) but settled on the short form BB for performance and stability. When applied to gradient descent, short form BB sets the stepsizes according to $\tau^{(k)} = p^\top q/(q^\top q)$, where $p = x^{(k)} - x^{(k-1)}$ and $q = \tilde{\nabla}^{(k)} - \tilde{\nabla}^{(k-1)}$. When applied to SVRG these values are computed in iterations when the full gradient is recomputed.

2.3 Preconditioning

Preconditioners are essential for accelerating iterative reconstruction algorithms by stabilizing admissible stepsize and adapting them to individual components of the solution. Effectively, image

components with large gradient variance get smaller updates, and vice versa. This can have a dramatic effect in PET image reconstruction (and machine learning applications) due to wildly varying range of local Lipschitz constants. Motivated by Newton's method, many preconditioners aim to approximate the inverse of the Hessian and thus may allow unit stepsizes. However, computing full Hessians is impractical in high dimensions, motivating the need for efficient approximations.

Preconditioners based on only the data-fidelity are standard in PET. The most prominent example is

$$D_{\mathrm{MLEM}}(x) = \operatorname{diag}\left(\frac{x+\delta}{A^\top 1}\right),$$

which can be derived from the gradient descent interpretation of maximum likelihood expectation maximization (MLEM). Here, the division of the two vectors is interpreted componentwise. Since $x \geq 0$ and $A^\top 1 > 0$, a small constant $\delta > 0$ ensures that the every diagonal entry of the preconditioner is non-zero. D_{MLEM} tends to work well for weak priors (e.g., in low-noise scenarios). However, it often underperforms as it does not account for the strength of the prior. This can either jeopardize the convergence behavior or require significant stepsize tuning.

Let

$$D_{\beta\mathcal{S}}(x) = \operatorname{diag}\left(\frac{1}{\operatorname{diag}(H_{\beta\mathcal{S}}(x))}\right)$$

be the inverse of the diagonal of the Hessian of the regularizer. In this work we use diagonal preconditioners that combine the data-fidelity and the prior terms via the (scaled) harmonic mean between D_{MLEM} and $D_{\beta\mathcal{S}}$. For scalars $a, b > 0$, the harmonic mean is given by

$$h(a,b) = \frac{2}{\frac{1}{a}+\frac{1}{b}}.$$

Since our preconditioners are diagonal, this can be readily extended to define for some $\alpha > 0$

$$\begin{aligned} D(x) &= \frac{1}{2} h(D_{\mathrm{MLEM}}(x), \alpha^{-1} D_{\beta\mathcal{S}}(x)) \\ &= \left(D_{\mathrm{MLEM}}^{-1}(x) + \alpha D_{\beta\mathcal{S}}^{-1}(x)\right)^{-1} \\ &= \operatorname{diag}\left(\frac{x+\delta}{A^\top 1 + \alpha \operatorname{diag}(H_{\beta\mathcal{S}}(x))(x+\delta)}\right). \end{aligned} \tag{9}$$

Note that it satisfies $D(x) \leq \min\{D_{\mathrm{MLEM}}(x), \alpha^{-1} D_{\beta\mathcal{S}}(x)\}$. While this may look like an ad-hoc choice, if D_{MLEM} and $\alpha^{-1} D_{\beta\mathcal{S}}$ are good approximations to their respective Hessians, then the harmonic mean D will be a good approximation to Hessian of the entire smooth term $\mathcal{J}$.

We tested several alternatives to (9), such as taking an componentwise minimum between D_{MLEM} and $D_{\beta\mathcal{S}}$, reweighing their contributions, using the Kailath variant of the Woodbury identity (together with the diagonal approximation) to estimate the inverse of the Hessian, and other variants. The selected preconditioner provided the best compromise between computational costs required to compute it and algorithmic performance. Traditional second order methods update the preconditioner in every iteration, which is costly. Preconditioner (9) is much cheaper and, as experiments show, requires updating only in the first 3 epochs, after which it stabilizes with no performance gain from further updates.

2.4 Subset Selection and Sampling

Subset-based reconstruction algorithms enhance the convergence speed of traditional iterative methods by dividing the projection data into multiple subsets and performing updates using partial measurement data. While this approach can offer significant computational advantages, careful selection of the number of subsets is critical. Using too many subsets can introduce artifacts and amplify noise, especially when subsets lack sufficient angular coverage, and increases the variance between successive updates, which can compromise the stability and convergence properties. Conversely, selecting too few subsets diminishes the acceleration benefit and causes behavior similar to classical methods, such as MLEM, which are known for their slow convergence. The number of subsets n is typically chosen as a divisor of the total number of projection angles (or views), allowing the data to be partitioned evenly. Subsets are then constructed to ensure that each is representative and uniformly distributed. We found that using approximately 25 subsets provides

a good tradeoff between reconstruction quality and computational speed in most scenarios, given the current computational requirements and scanner configurations.

To determine the order in which subsets are accessed we consider the following standard choices

Herman–Meyer order [26] is a well-established deterministic choice that is based on the prime decomposition of the number of subsets.

Uniformly random with replacement is the most common choice in machine learning applications. In each iteration the subset index i is chosen by taking a sample from $\{1, \dots, n\}$ uniformly at random.

Uniformly random without replacement randomizes access to subset indices but ensures n successive iterates cycle through all the data by computing a permutation of $(1, \dots, n)$ in each epoch.

Importance sampling uses a weighted variant of uniform sampling with replacement. For each $1 \leq i \leq n$ we assign a probability $p_i \geq 0$, such that $\sum_{i=1}^{n} p_i = 1$. When Lipschitz constants L_i are known then $p_i = L_i / \sum_{j=1}^{n} L_j$ is a common choice.

Since Lipschitz constants L_i are unknown in PET, we propose an alternative importance sampling strategy for SVRG. Namely, when the full gradient estimator is updated we compute $p_i = \|\nabla \mathcal{J}_i(x)\| / \sum_{j=1}^{n} \|\nabla \mathcal{J}_j(x)\|$, where x is the current image estimate. This incurs minimal computational overhead, since in SVRG all the subset gradients are already recomputed.

Lastly, drawing inspiration from the Herman–Meyer ordering, which is designed to maximize information gain between successive updates, and incorporating the concept of random sampling without replacement to ensure full coverage of subsets in each epoch with varying order, we propose the following novel subset ordering strategy.

Cofactor order begins by identifying all generators of the cyclic group associated with the number of subsets, n, which are identified as positive integers $k < n$ that are coprime with n, meaning they share no common prime factors with it. These generators are then ranked according to their proximity to two reference points: $0.3n$ and $0.7n$, aiming to balance spread and randomness. In each epoch, the next available generator from this sorted list is selected and used to define a new traversal of the cyclic group, thereby determining the order in which subsets are accessed (i.e., one subset index per iteration). Once the list of generators is exhausted, it is reinitialized, and the process repeats for subsequent epochs.

3 Numerical Simulation Experiments

To validate and refine the algorithmic components introduced in the previous section, we conducted a comprehensive suite of fast *inverse-crime* simulations. By simulating a simplified yet realistic PET scanner using the pure GPU mode of `parallelproj v1.10.1` [27], iterative reconstructions could be run in seconds. This enabled a systematic exploration of the effects of various factors on convergence behavior, including the choice of stochastic algorithm, preconditioner, step-size strategy, number of subsets, subset sampling method, time-of-flight (ToF) versus non-ToF data, count levels, and regularization strength.

3.1 Simulation Setup

All experiments used a simulated cylindrical (polygonal) scanner with a diameter of 600 mm and a length of 80 mm, composed of 17 rings with 36 modules each (12 detectors per module). Simulated ToF resolution was 390 ps, and a 4 mm isotropic Gaussian kernel in image space was used to model limited spatial resolution. Emission data was binned into a span 1 sinogram (289 planes, 216 views, 353 radial bins, 25 ToF bins). A simple 3D elliptical phantom was forward-projected (accounting for water attenuation), contaminated with a smooth background sinogram, and corrupted by Poisson noise to simulate realistic emission data. Low and high count regimes were simulated with 10^7 and 10^8 true events, respectively. Reconstruction was performed at image size $161 \times 161 \times 33$ voxels with a 2.5 mm isotropic spacing.

Reference reconstructions (see Fig. 1) were obtained by running 500 iterations of preconditioned L-BFGS-B with three relative regularization strengths $\tilde{\beta} \in \{1, 4, 16\}$. The regularization parameter β was scaled as

$$\beta = \tilde{\beta} \times 2 \times 10^{-4} \times \frac{\text{true counts}}{3 \times 10^7}\,. \tag{10}$$

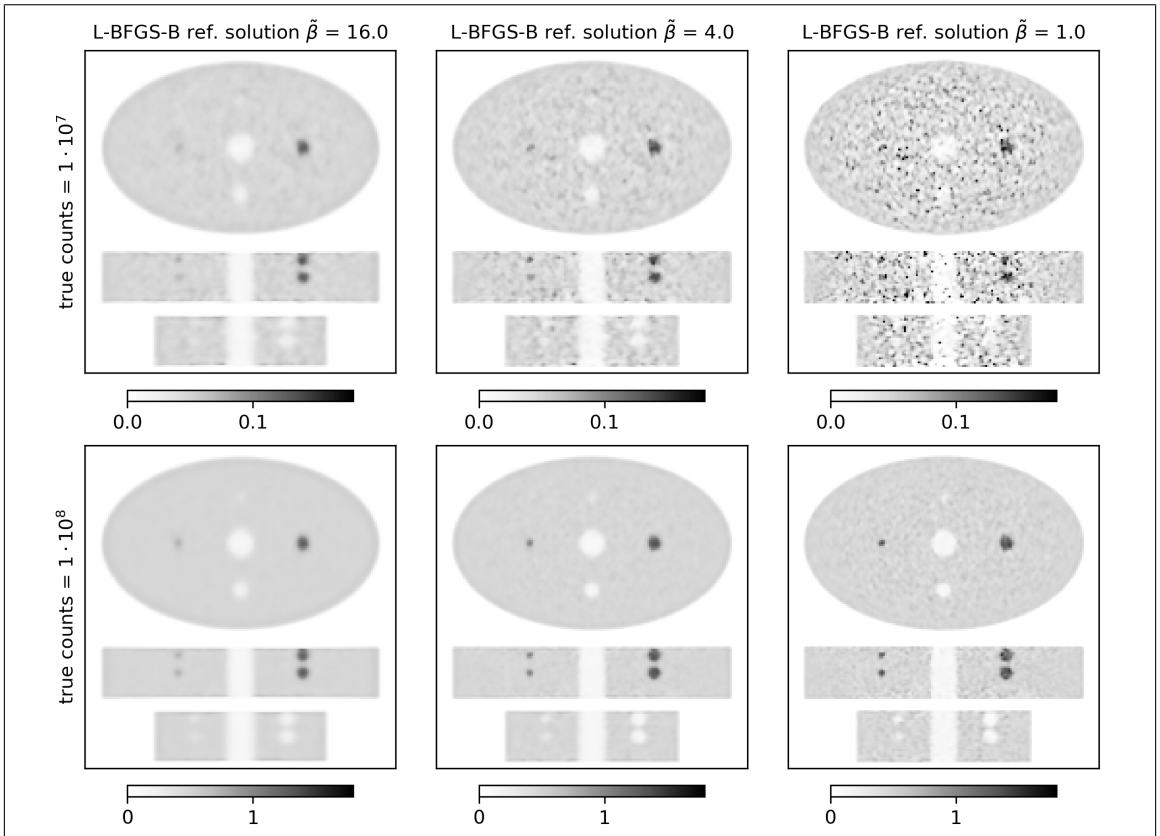


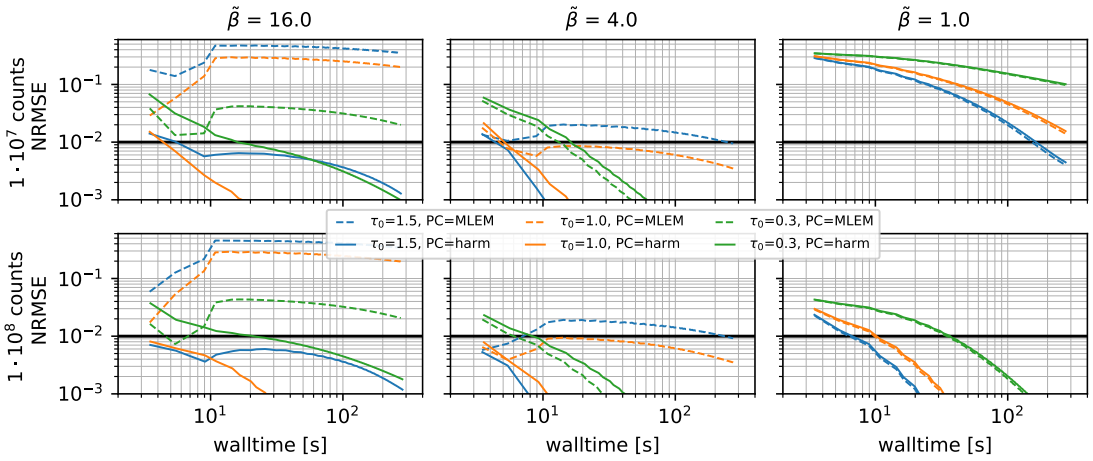
Figure 1: Stacked central transversal, coronal and sagittal slices of L-BFGS-B reference reconstructions of the ellipse phantom. Each column shows a different level of regularization ($\tilde{\beta}$) increasing from left to right. Top row shows results for 10^7 true counts, and bottom row for 10^8 true counts.

This ensures that reconstructions with the same $\tilde{\beta}$ at different count levels show comparable resolution. All stochastic reconstructions were initialized with one epoch of OSEM (with 27 subsets). Convergence was measured by the normalized root mean square error (NRMSE) excluding cold background around the elliptical phantom, normalized by the intensity of the largest background ellipsoid. In line with the NRMSE target threshold used in the PETRIC challenge, we consider the point where NRMSE was less than 0.01 as a marker of practical convergence. The data was divided into n subsets by selecting every n-th view. Unless stated otherwise, in each epoch subsets were drawn uniformly at random without replacement. All runs were performed using an NVIDIA RTX A4500 GPU. The code for all our simulation experiments as well as our submissions to PETRIC is available on GitHub.

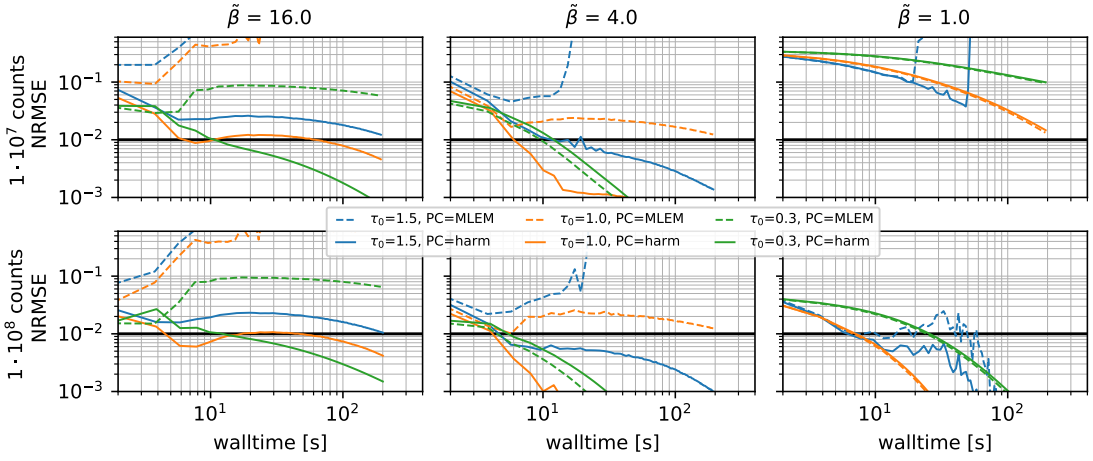
3.2 Main Simulation Results

Algorithm and preconditioner effects (see Fig. 2): When comparing SVRG, SAGA and plain SGD under a vanishing stepsize schedule $\tau^{(k)} = \tau^{(0)}/(1 + 0.02\,k/n)$ with $\tau^{(0)} \in \{0.3, 1.0, 1.5\}$ and $n = 27$, we made the following observations.

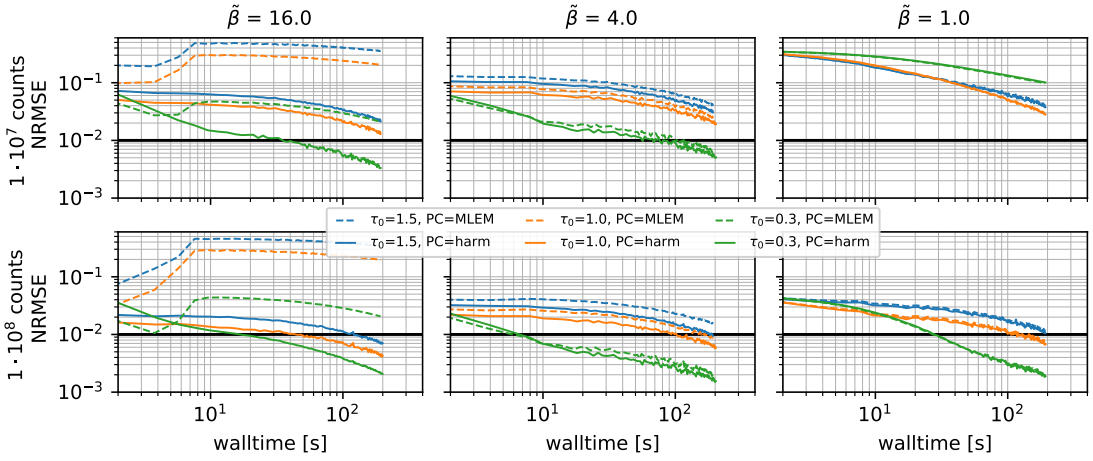
- SVRG and SAGA consistently outperform SGD in all count and regularization regimes.
- The harmonic-mean preconditioner (9) is crucial: under strong regularization $\tilde{\beta} = 16$, the classic MLEM preconditioner diverges or converges extremely slowly (depending on the chosen stepsize), whereas the harmonic-mean variant converges reliably in every scenario.
- SVRG with the harmonic preconditioner, $\tau^{(0)} = 1$ and $\eta = 0.02$ (giving mild decay) yields the fastest convergence for medium and high $\tilde{\beta}$. For low regularization, a slightly larger $\tau^{(0)}$ (up to 1.5 or 2.5) can accelerate convergence.
- Across all methods, convergence was slower in the case of low regularization $\tilde{\beta} = 1$.



(a) SVRG

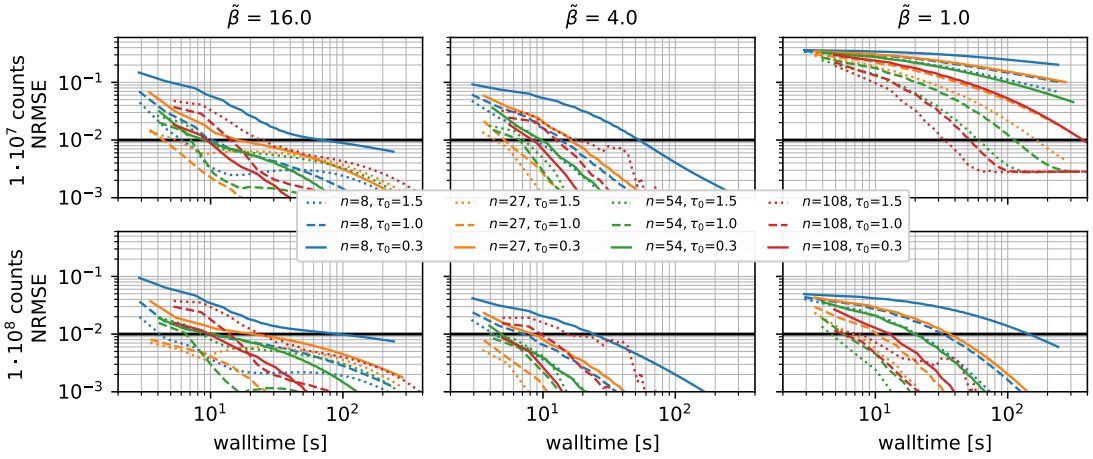


(b) SAGA

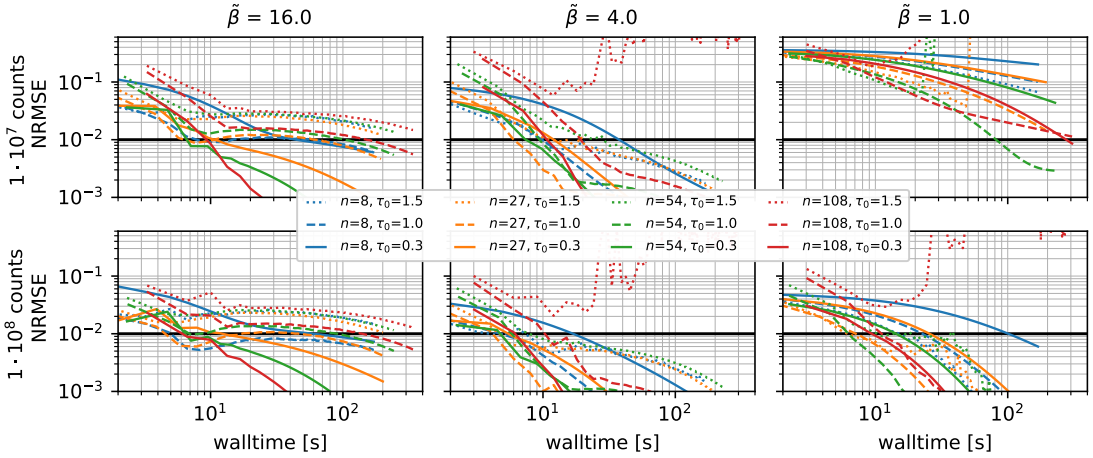


(c) SGD

Figure 2: Reconstruction performance in terms of NRMSE versus walltime for **SVRG, SAGA, SGD**, for MLEM (dashed lines) and harmonic (solid lines) **preconditioners** (PC) and three **initial stepsizes** ($\tau^{(0)}$) represented by different colors, using 27 subsets, a gentle stepsize decay with $\eta = 0.02$, 100 epochs, and subset selection without replacement. Results are shown for three levels of regularization ($\tilde{\beta}$) and two count levels. **Note the logarithmic scale on the x and y axes.** For each combination of n and $\tau^{(0)}$, the outcome of **1 run** is displayed. The thick horizontal black line shows the NRMSE target threshold of 10^{-2} used in PETRIC.



(a) SVRG



(b) SAGA

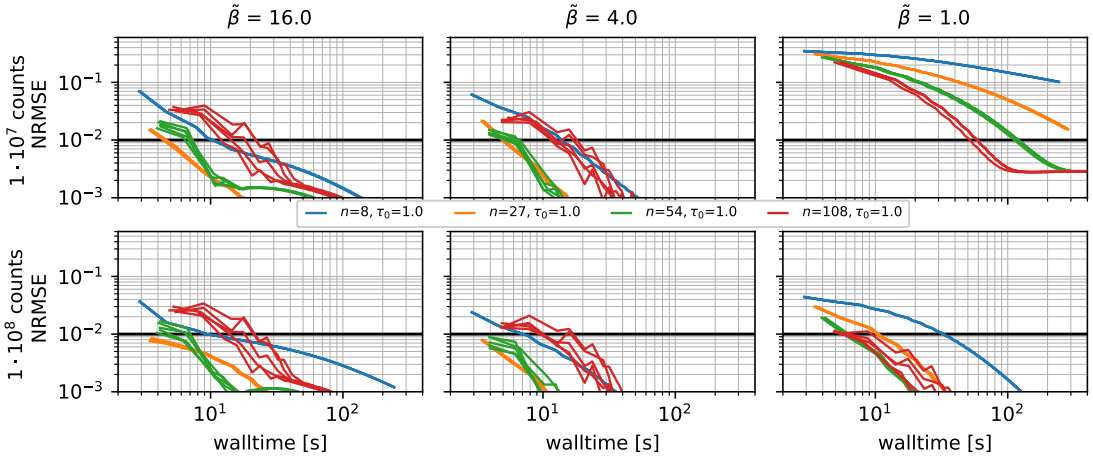
Figure 3: Performance in terms of NRMSE versus walltime for **SVRG and SAGA**, for different **number of subsets** n and **initial stepsizes** $\tau^{(0)}$, using the **harmonic preconditioner**, a gentle stepsize decay with $\eta = 0.02$, 100 epochs, and subset selection without replacement. Results are shown for three levels of regularization $\tilde{\beta}$ and two count levels. For each combination of n and $\tau^{(0)}$, the outcome of **1 run** is displayed. The thick horizontal black line shows the NRMSE target threshold of 10^{-2} used in the PETRIC challenge.

Impact of the number of subsets (see Fig. 3): Fixing the harmonic preconditioner and vanishing stepsize rule $\tau^{(0)} = 1, \eta = 0.02$, we varied the number of subsets $n \in \{8, 27, 54, 108\}$:

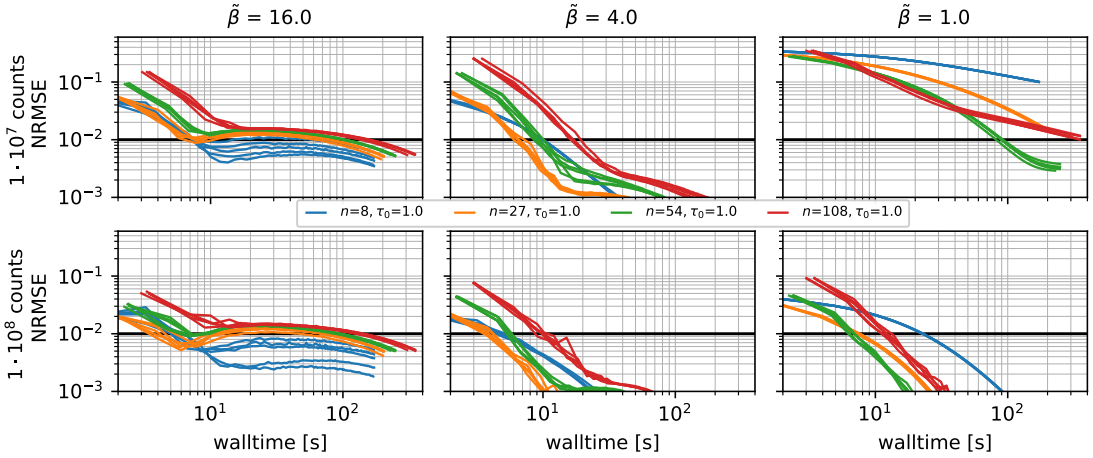
- SVRG achieves optimal walltime convergence at $n = 27$ under medium to high $\tilde{\beta}$. Lower $\tilde{\beta}$ benefit from using a greater number of subsets.
- Optimal values of n and $\tau^{(0)}$ for SAGA depend strongly on $\tilde{\beta}$: high $\tilde{\beta}$ favors a larger number of subsets with smaller $\tau^{(0)}$, medium $\tilde{\beta}$ favors $n = 27$ with $\tau^{(0)} \approx 1$, and low $\tilde{\beta}$ favors $n \approx 54$.
- Overall, SVRG with optimized settings achieves faster convergence compared to SAGA with optimized settings.

Stability across repeated runs using different subsets orders (see Fig. 4): We run five independent runs (changing the random seed used for the random subset selection) of the reconstructions using SVRG, the harmonic preconditioner, $\tau^{(0)} = 1$, $\eta = 0.02$ and $n \in \{8, 27, 54, 108\}$. The run-to-run NRMSE variation is small, especially at $n = 27$, confirming low variance introduced by the stochastic subset selection in this setting.

Subset sampling strategy (see Fig. 5): Comparing Herman–Meyer order, uniform sampling at random with and without replacement, importance sampling, and cofactor strategies for selecting the order of subsets for SVRG with $\tau^{(0)} = 1$, $n = 27$, $\eta = 0.02$, we observe negligible differences



(a) SVRG



(b) SAGA

Figure 4: Same as Fig. 3 showing the results of **5 runs** a different random seed for the subset selection.

between all subset selection rules in simulated scenarios, with some minor benefits for sampling without replacement and cofactor sampling.

Stepsize rules (see Fig. 6). We see that for SVRG, $n = 27$, and the harmonic preconditioner:

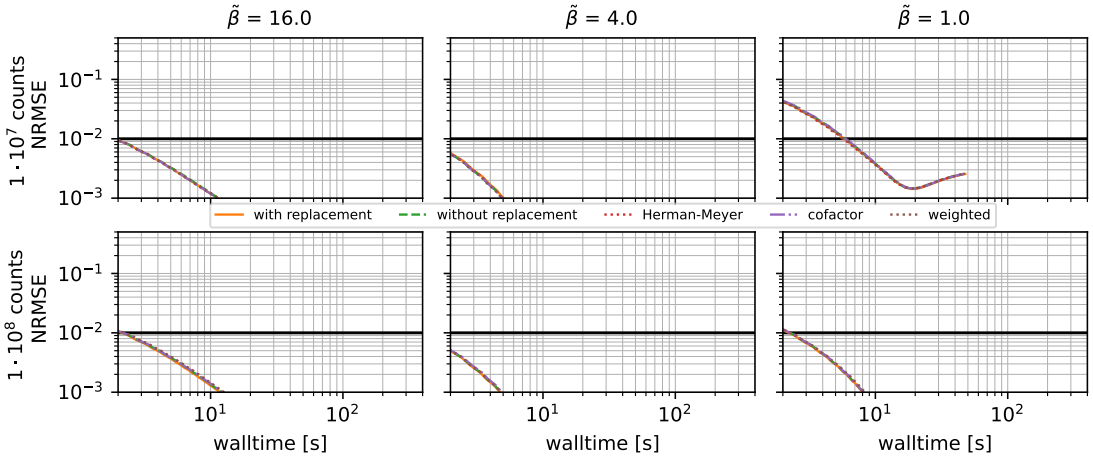
- At low $\tilde{\beta}$, adaptive rules (short-form BB or heuristic ALG1) modestly outperform a simple decay.
- However, in the medium-to-high $\tilde{\beta}$ regime, a constant or decaying initialization $\tau^{(0)} = 1$ yields superior ToF reconstruction performance compared to adaptive BB schemes.

3.3 Simulation-derived Conclusions

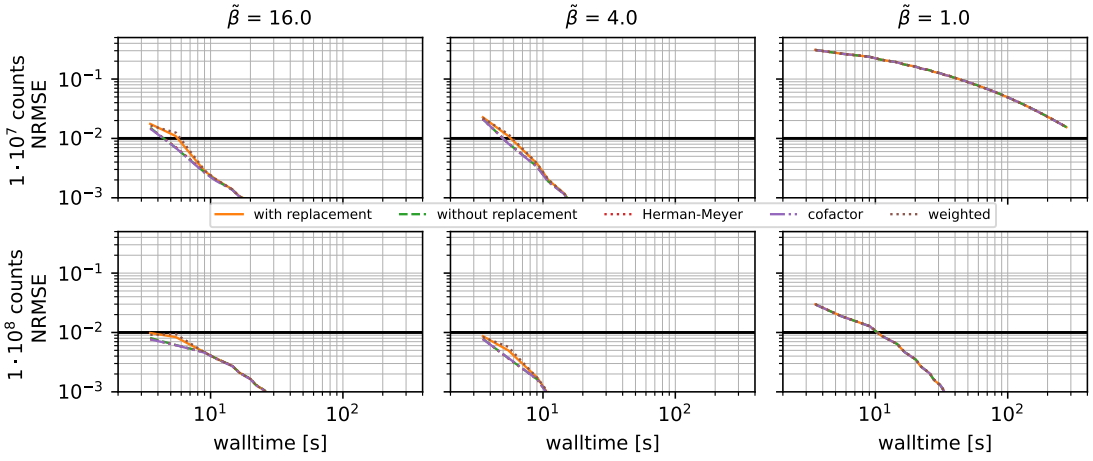
The inverse-crime simulation study motivated the design of our algorithms submitted to the PET-RIC challenge in the following way:

- The **harmonic-mean preconditioner** was essential to achieve stable convergence with $\tau^{(0)} \approx 1$ across count and regularization regimes.
- **SVRG** slightly outperformed SAGA in robustness and speed, and both outperformed SGD.
- A moderate number of subsets, $n \approx 27$, leads to the fastest convergence times.

These guidelines directly informed our implementation choices for the three submitted algorithms explained in detail in the next section.

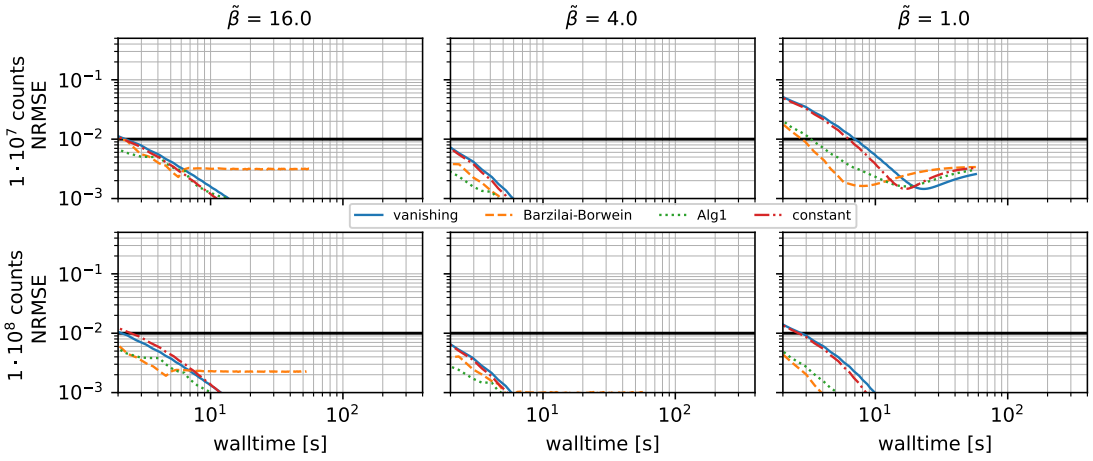


(a) non-ToF reconstructions

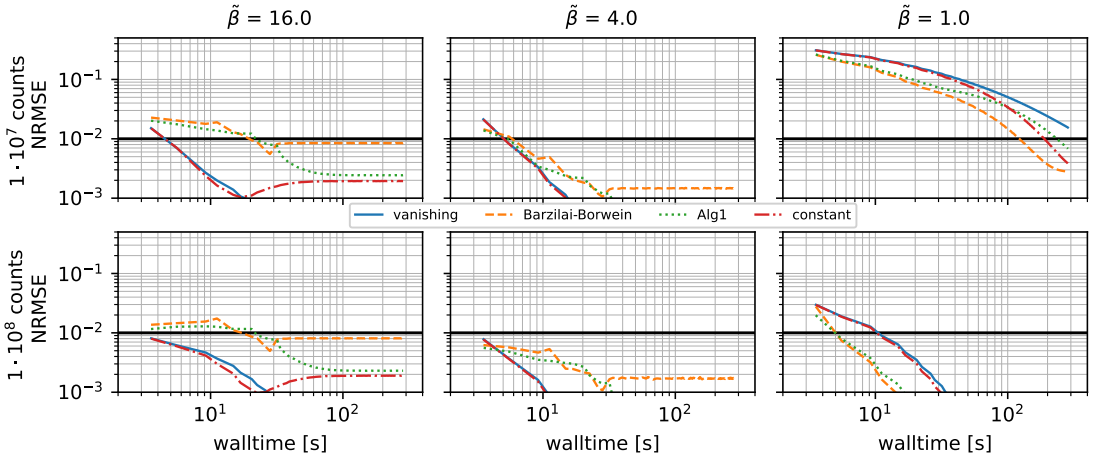


(b) ToF reconstructions

Figure 5: Same as Fig. 3 (SVRG only) showing the results for different subset sampling strategies, $n = 27$ subsets, the harmonic preconditioner, an initial stepsize $\tau^{(0)} = 1$ and gentle stepsize decay using $\eta = 0.02$ for non-ToF (top) and ToF reconstructions (bottom).



(a) non-ToF reconstructions



(b) ToF reconstructions

Figure 6: Same as Fig. 5 showing the results for different stepsize strategies, 27 subsets, and the harmonic preconditioner for non-ToF (top) and ToF reconstructions (bottom).

Table 1: Key hyperparameters of the three submitted algorithms

	ALG1	**ALG2**	**ALG3**
Gradient estimator	SVRG	same as ALG1	same as ALG1
Preconditioner	Harmonic mean	same as ALG1	same as ALG1
Preconditioner update epochs	1, 2, 3	1, 2, 4, 6	1, 2, 4, 6
Number of subsets	Divisor of number of views closest to 25	same as ALG1	Divisor of number of views closest to 24.2
Subset selection rule	fixed random sequence without replacement	same as ALG1	cofactor
Stepsize rule	$\begin{cases} 3 & k < 10 \\ 2 & 10 \le k < 100 \\ 1.5 & 100 \le k < 200 \\ 1 & 200 \le k < 300 \\ 0.5 & 300 \le k \end{cases}$	$\begin{cases} \min(\tau_{\text{bb}}^{(k)}, 3) & k < 10 \\ \min(\tau_{\text{bb}}^{(k)}, 2.2) & 10 \le k < 2n \\ \min(\tau_{\text{bb}}^{(k)}, 1) & 2n \le k \end{cases}$ with $\tau_{\text{bb}}^{(k)}$ the short BB step calculated at the end of epochs 2, 4, and 6.	same as ALG2

4 Submitted Algorithms and Their Performance

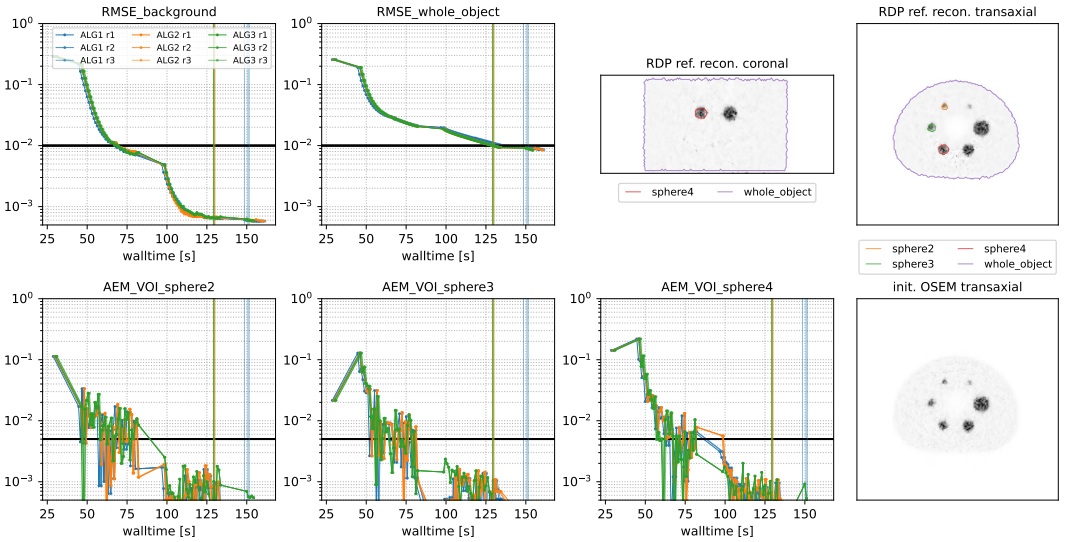
Based on the insights gained from the inverse-crime simulations in the previous section, we implemented and submitted three closely related algorithms (termed **ALG1**, **ALG2**, and **ALG3**) to the PETRIC challenge under the team name MaGeZ. All three algorithm use SVRG as the underlying stochastic gradient algorithm and apply the harmonic-mean preconditioner (9). Pseudo-code that forms the basis of all three algorithms is given in Algorithm 1 in Appendix A.2.

The available PETRIC training datasets were primarily used to fine-tune the algorithm hyperparameters, namely (i) number of subsets, (ii) subset selection strategy, (iii) stepsize rule and (iv) update-frequency of the preconditioner. These are the only distinguishing features among the submitted algorithms and our choices are summarized in Table 1. ALG1 and ALG2 use the number of subsets as the divisor of the number of view closest to 25. ALG3 further modifies the subset count slightly using the divisor closest to 24.2 (with the goal of selecting a smaller number of subsets in some of the training datasets). In ALG1 and ALG2 subsets are chosen uniformly at random without replacement in each iteration of each epoch. ALG3 uses the proposed cofactor rule. ALG1 updates the preconditioner at the start of epochs 1, 2, and 3. ALG2 and ALG3 update the preconditioner at the start of epochs 1, 2, 4, and 6. ALG1 uses a fixed, piecewise stepsize schedule, while ALG2 and ALG3 employ a short BB rule for adaptive stepsize reduction, which is computed at the start of epochs 1, 2, 4, and 6.

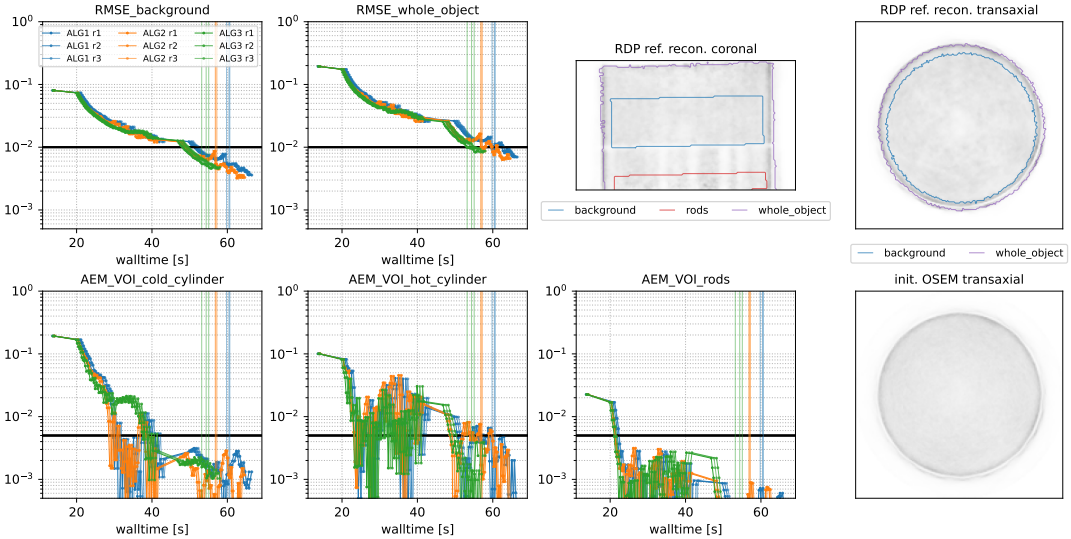
4.1 Performance on PETRIC Test Datasets

Figures 7 and 8 present the convergence behavior of all three submitted algorithms in terms of whole-object NRMSE, background NRMSE, and multiple volume-of-interest (VOI) mean absolute error metrics (AEM). Each dataset was reconstructed three times with all three algorithms using a local an NVIDIA RTX A4500 GPU. From the two figures, we observe:

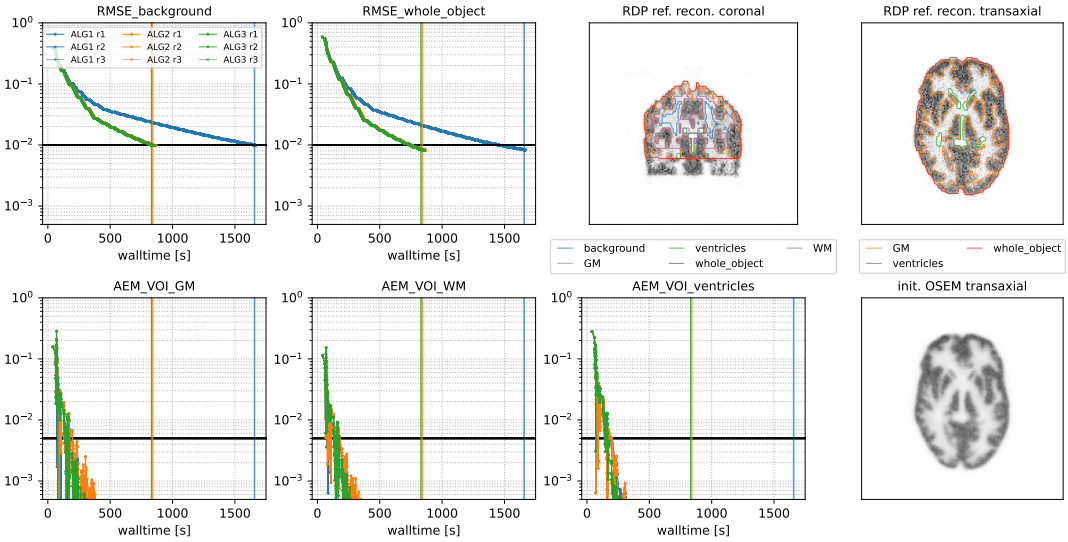
- **All algorithms converge** reliably across all datasets and runs.
- **ALG2 and ALG3 perform similarly**, and both slightly outperform ALG1 in most cases. In the Vision600 Hoffman dataset, ALG1 almost takes twice along to reach the convergence threshold compared to ALG2 and ALG3.
- **For the DMI4 NEMA, NeuroLF Esser, and Mediso low-count datasets**, convergence is reached very quickly both in terms of walltime and epoch count, typically within 4 epochs.
- **Vision600 Hoffman dataset** shows the slowest convergence, requiring more than 23 epochs (594 updates) for ALG2 and ALG3, and more than 47 epochs (1184 updates) for ALG1.



(a) DMI4 NEMA test dataset

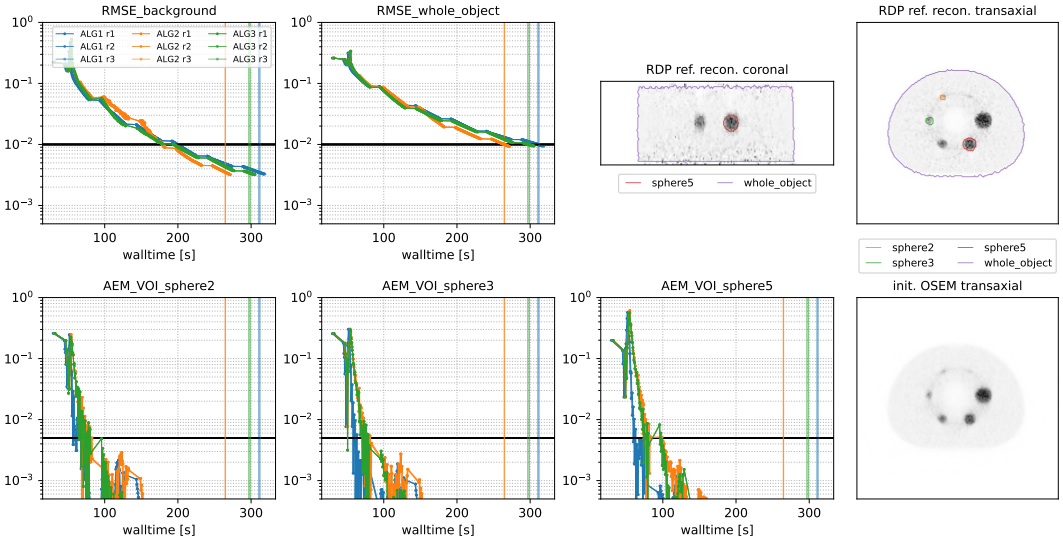


(b) NeuroLF Esser test dataset

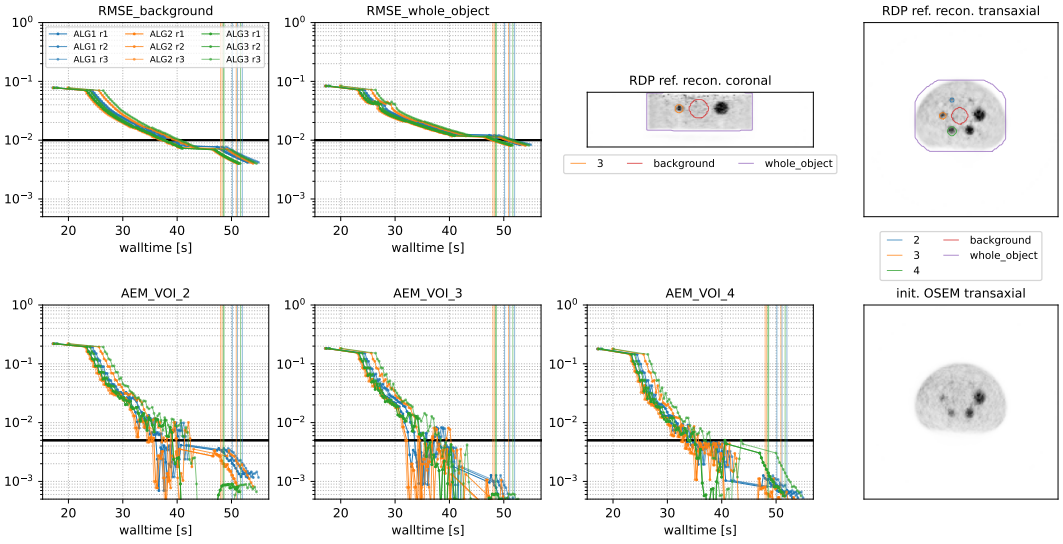


(c) Vision600 Hoffman test dataset

Figure 7: Performance metrics of our three submitted algorithms evaluated on 3 representative PETRIC test datasets using 3 repeated runs. The vertical lines indicate the time when the threshold of all metrics were reached. Note the logarithmic scale on the y-axis and the linear scale on the x-axis. The top right images show a coronal and transaxial slice of the reference reconstruction alongside contour lines of the volumes of interest used for the metrics. The bottom right image shows the same transaxial slice of the OSEM reconstruction used for initialization of all algorithms.



(a) D690 NEMA test dataset



(b) Mediso low count test dataset

Figure 8: Same as Fig. 7 for 2 more datasets.

- **Inter-run variability** is low; timing differences between runs are within 1–2 seconds.
- Across all datasets, **whole-object NRMSE is the slowest metric to converge**, becoming the bottleneck in determining the final convergence time.

Closer inspection of the stepsize behavior on the Vision600 Hoffman dataset reveals that the slower convergence of ALG1 is due to its lower final stepsize, implemented as a "safety feature". After 300 updates, ALG1 reduces $\tau^{(k)}$ to 0.5, whereas ALG2 and ALG3 continue to use $\tau^{(k)} = 1.0$, since the BB-based calculated adaptive stepsizes were bigger in this dataset. This difference explains the kink observed in ALG1's convergence curves around 450 s.

5 Discussion

We now want to turn to a discussion on what we believe are important and interesting aspects of this work.

In our view, by far the most important feature of our algorithms is the improved preconditioner taking into account Hessian information of the regularizer. This meant that the stepsize choices generalized better across a range of scanners, objects, noise levels and regularization strengths. We settled on SVRG for our gradient estimator, this choice is not as clear and may be different for other variants of the reconstruction problem. In our experience, employing a sophisticated method to control variance is important, but the specific approach used (e.g., SVRG or SAGA) appears to be less critical. In contrast, other factors like stepsizes and sampling strategies seem to have a relatively minor impact, as the algorithms are not particularly sensitive to these choices.

A key aspect in our approach was to consider what can be effectively computed and what cannot. For the RDP it is easy to compute the gradient and the diagonal Hessian, but other operations such as the proximity operator or the full Hessian are much more costly. Similarly, the ideal number of subsets is largely a computational efficiency question. It has been observed numerous times that theoretically fewer epochs are needed with larger number of subsets. However, practically this means that the overhead per epoch increases, e.g., as the gradient is computed in each iteration of the epoch. These have the be traded off against each other.

Speaking of the RDP, we noticed a couple of interesting features which we have not exploited in our work. First, the diagonal Hessian of the RDP is very large in the background where the activity is small. Second, while its gradient has a Lipschitz constant, similar to the total variation and its smoothed variants, algorithms which do not rely on gradients might be beneficial.

Between the three algorithms ALG2 and ALG3 consistently performed either similar or better than ALG1. Comparing them to the submissions of other teams it is worth noting that for almost all datasets, they perform far better than any of the other competitors which lead to MaGeZ winning the challenge overall [28].

Coordination between simulation insights and algorithm design was essential to our approach. Local testing allowed us to validate generalization before submission. Across datasets, we favored robustness over aggressive tuning. Refinement came from iterative testing rather than theoretical guarantees alone. Above all, our goal was to develop an algorithm that performs well out-of-the-box.

6 Conclusions

In this paper we presented our strategy and thought process behind designing the winning strategy for the 2024 PETRIC challenge. We identified the key parameters for PET image reconstruction algorithms via realistic yet very fast simulations. The harmonic mean preconditioner helped to overcome the biggest roadblock of the challenge which was the tuning of parameters for a variety of settings with various scanner models, phantoms and regularization strengths.

Methods and Materials

We do not use results from animal or human subject research. We use computer simulated data, with simulated scanners and measurements for the majority of the results, which can be found on zenodo. Also, publicly available PET data provided by PETRIC is used.

Acknowledgments

We acknowledge support from the EPSRC: MJE (EP/Y037286/1, EP/S026045/1, EP/T026693/1, EP/V026259/1) and ZK (EP/X010740/1). GS acknowledges the support from NIH grant R01EB029306 and FWO project G062220N.

A Appendix

A.1 Gradient and Hessian of the RDP

For completeness, we present here the first and second derivatives of the RDP (3), i.e., the gradient and the diagonal of the Hessian. Both of these are used in our proposed solution.

Let $d_{i,j} = x_i - x_j$, $s_{i,j} = x_i + x_j$ and $\phi_{i,j} = s_{i,j} + \gamma|d_{i,j}| + \varepsilon$. Then the first derivative is given by

$$\partial_{x_i}\mathcal{S}(x) = \sum_{j\in N_i} w_{i,j}\kappa_i\kappa_j \frac{d_{i,j}(2\phi_{i,j} - (d_{i,j} + \gamma|d_{i,j}|))}{\phi_{i,j}^2},$$

and the second by

$$\partial_{x_i}^2\mathcal{S}(x) = 2\sum_{j\in N_i} w_{i,j}\kappa_i\kappa_j \frac{(s_{i,j} - d_{i,j} + \varepsilon)^2}{\phi_{i,j}^3}.$$

A.2 Pseudocode for submitted preconditioned SVRG algorithm

Algorithm 1 Preconditioned SVRG Algorithm

Require: initial image: x, number of subsets: n, stepsize rule: `stepsize`, sampling rule: `subset`, diagonal preconditioner rule: `preconditioner`, list of iterations to update the preconditioner: `update_pc_iters`, update gradient at anchor point every ω epochs (default=2)

for $k = 0, 1, \ldots$ **do**
  if $k \in$ `update_pc_iters` **then**
    $D \leftarrow$ `preconditioner`(x) ▷ update preconditioner via (9)
  end if
  if $k \bmod (\omega n) = 0$ **then**
    for $i = 1$ **to** n **do**
      $\hat{g}_i \leftarrow \nabla\mathcal{J}_i(x)$ ▷ calculate all subset gradients at snapshot image
    end for
    $\hat{g} \leftarrow \sum_{i=1}^{n} \hat{g}_i$
    $\tilde{\nabla} \leftarrow \hat{g}$
  else
    $i \leftarrow$ `subset`(k)
    $\tilde{\nabla} \leftarrow n\big(\nabla\mathcal{J}_i(x) - \hat{g}_i\big) + \hat{g}$
  end if
  $\tau \leftarrow$ `stepsize`(k)
  $x \leftarrow x - \tau D\tilde{\nabla}$
  if stopping criterion is reached **then return** x
  end if
end for

Acronyms

BB Barzilai–Borwein. 4, 10, 13, 16

KL Kullback–Leibler. 1, 2

MLEM maximum likelihood expectation maximization. 5, 7, 8

NRMSE normalized root mean square error. 7, 8, 9, 13, 16

OSEM ordered subsets expectation maximization. 1, 7, 14

PET positron emission tomography. 1, 2, 3, 4, 5, 6

PETRIC PET Rapid Image Reconstruction Challenge. 1, 2, 3, 7, 8, 9, 10, 13, 14, 16

RDP relative difference prior. 1, 2, 16, 17

SAGA Stochastic Averaged Gradient Amelioré. 1, 3, 4, 7, 8, 9, 10, 16

SGD Stochastic Gradient Descent. 1, 3, 4, 7, 8, 10

SIRF Synergistic Image Reconstruction Framework. 2

SVRG Stochastic Variance Reduced Gradient. 1, 4, 6, 7, 8, 9, 10, 11, 13, 16, 17

ToF time-of-flight. 6, 10, 11, 12

References

[1] Johan Nuyts, Dirk Bequé, Patrick Dupont, and Luc Mortelmans. A concave prior penalizing relative differences for maximum-a-posteriori reconstruction in emission tomography. *IEEE Transactions on nuclear science*, 49(1):56–60, 2002.

[2] Casper da Costa-Luis, Matthias J. Ehrhardt, Christoph Kolbitsch, Evgueni Ovtchinnikov, Edoardo Pasca, Kris Thielemans, and Charalampos Tsoumpas. Petric: Pet rapid image reconstruction challenge, 2025.

[3] Eugene J. Teoh, Daniel R. McGowan, Ruth E. Macpherson, Kevin M. Bradley, and Fergus V. Gleeson. Phantom and clinical evaluation of the bayesian penalized likelihood reconstruction algorithm q.clear on an lyso pet/ct system. *Journal of Nuclear Medicine*, 56(9):1447–1452, July 2015.

[4] Eugene J. Teoh, Daniel R. McGowan, Kevin M. Bradley, Elizabeth Belcher, Edward Black, and Fergus V. Gleeson. Novel penalised likelihood reconstruction of pet in the assessment of histologically verified small pulmonary nodules. *European Radiology*, 26(2):576–584, May 2015.

[5] Sangtae Ahn, Steven G Ross, Evren Asma, Jun Miao, Xiao Jin, Lishui Cheng, Scott D Wollenweber, and Ravindra M Manjeshwar. Quantitative comparison of osem and penalized likelihood image reconstruction using relative difference penalties for clinical pet. *Physics in Medicine and Biology*, 60(15):5733–5751, July 2015.

[6] Sangtae Ahn and Jeffrey A Fessler. Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms. *IEEE transactions on medical imaging*, 22(5):613–626, 2003.

[7] Robert Twyman, Simon Arridge, Zeljko Kereta, Bangti Jin, Ludovica Brusaferri, Sangtae Ahn, Charles W Stearns, Brian F Hutton, Irene A Burger, Fotis Kotasidis, et al. An investigation of stochastic variance reduction algorithms for relative difference penalized 3d pet image reconstruction. *IEEE Transactions on Medical Imaging*, 42(1):29–41, 2022.

[8] Evgueni Ovtchinnikov, Richard Brown, Christoph Kolbitsch, Edoardo Pasca, Casper da Costa-Luis, Ashley G Gillman, Benjamin A Thomas, Nikos Efthimiou, Johannes Mayer, Palak Wadhwa, et al. Sirf: synergistic image reconstruction framework. *Computer Physics Communications*, 249:107087, 2020.

[9] H. Malcolm Hudson and Richard S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE transactions on medical imaging*, 13(4):601–609, 1994.

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

[11] Željko Kereta, Robert Twyman, Simon Arridge, Kris Thielemans, and Bangti Jin. Stochastic em methods with variance reduction for penalised pet reconstructions. *Inverse Problems*, 37(11):115006, 2021.

[12] Georg Schramm and Martin Holler. Fast and memory-efficient reconstruction of sparse poisson data in listmode with non-smooth priors with application to time-of-flight pet. *Physics in Medicine & Biology*, 67(15):155020, 2022.

[13] Matthias Joachim Ehrhardt, Zeljko Kereta, Jingwei Liang, and Junqi Tang. A guide to stochastic optimisation for large-scale inverse problems. *Inverse Problems*, 2024.

[14] Evangelos Papoutsellis, Casper da Costa-Luis, Daniel Deidda, Claire Delplancke, Margaret Duff, Gemma Fardell, Ashley Gillman, Jakob S Jørgensen, Zeljko Kereta, Evgueni Ovtchinnikov, et al. Stochastic optimisation framework using the core imaging library and synergistic image reconstruction framework for pet reconstruction. *arXiv preprint arXiv:2406.15159*, 2024.

[15] Kristian Bredies, Karl Kunisch, and Thomas Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[16] Florian Knoll, Martin Holler, Thomas Koesters, Ricardo Otazo, Kristian Bredies, and Daniel K Sodickson. Joint mr-pet reconstruction using a multi-channel image regularizer. *IEEE transactions on medical imaging*, 36(1):1–16, 2016.

[17] Matthias J Ehrhardt, Pawel Markiewicz, Maria Liljeroth, Anna Barnes, Ville Kolehmainen, John S Duncan, Luis Pizarro, David Atkinson, Brian F Hutton, Sebastien Ourselin, et al. Pet reconstruction with an anatomical mri prior using parallel level sets. *IEEE transactions on medical imaging*, 35(9):2189–2199, 2016.

[18] Zhaoheng Xie, Reheman Baikejiang, Tiantian Li, Xuezhu Zhang, Kuang Gong, Mengxi Zhang, Wenyuan Qi, Evren Asma, and Jinyi Qi. Generative adversarial network based regularized image reconstruction for pet. *Physics in Medicine & Biology*, 65(12):125016, 2020.

[19] Imraj RD Singh, Alexander Denker, Riccardo Barbano, Željko Kereta, Bangti Jin, Kris Thielemans, Peter Maass, and Simon Arridge. Score-based generative models for pet image reconstruction. *Machine Learning for Biomedical Imaging*, 2:547–585, 2024.

[20] Yu-Jung Tsai, Georg Schramm, Sangtae Ahn, Alexandre Bousse, Simon Arridge, Johan Nuyts, Brian F Hutton, Charles W Stearns, and Kris Thielemans. Benefits of using a spatially-variant penalty strength with anatomical priors in pet reconstruction. *IEEE Transactions on Medical Imaging*, 39(1):11–22, 2019.

[21] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, volume 2, pages 1646–1654, 2014.

[22] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26, 2013.

[23] Conghui Tan, Shiqian Ma, Yu-Hong Dai, and Yuqiu Qian. Barzilai-borwein step size for stochastic gradient descent. *Advances in neural information processing systems*, 29, 2016.

[24] Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pages 14465–14499. PMLR, 2023.

[25] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019.

[26] Gabor T. Herman and Lorraine B. Meyer. Algebraic reconstruction techniques can be made computationally efficient (positron emission tomography application). *IEEE Transactions on Medical Imaging*, 12(3):600–609, 1993.

[27] Georg Schramm and Kris Thielemans. PARALLELPROJ–an open-source framework for fast calculation of projections in tomography. *Frontiers in Nuclear Medicine*, Volume 3 - 2023, 2024.

[28] Casper da Costa-Luis, Matthias J. Ehrhardt, Christoph Kolbitsch, Evgueni Ovtchinnikov, Edoardo Pasca, Kris Thielemans, and Charalampos Tsoumpas. Petric: Pet rapid image reconstruction challenge - leaderboard, 2025.