Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

ACCEPTED MANUSCRIPT • OPEN ACCESS

On Krylov methods for large-scale CBCT reconstruction

To cite this article before publication: Malena Sabaté Landman *et al* 2023 *Phys. Med. Biol.* in press https://doi.org/10.1088/1361-6560/acd616

On Krylov Methods for Large-Scale CBCT Reconstruction

Malena Sabaté Landman[1*], Ander Biguri[1*], Sepideh Hatamikia[2,3], Richard Boardman[4], John Aston[5], Carola-Bibiane Schönlieb[1]

[1]Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, Cambridge, UK
[2]Research center for Medical Image Analysis and Artificial Intelligence (MIAAI), Department of Medicine, Danube Private University, Krems, Austria
[3]Austrian Center for Medical Innovation and Technology (ACMIT), Wiener Neustadt, Austria
[4]µ-Vis X-ray Imaging Laboratory, University of Southampton, Southampton, UK
[5]Department of Pure Mathematics and Mathematical Statistics (DPMMS), University of Cambridge, Cambridge, UK
[*]These authors contributed equally to this work

E-mail: m.sabate.landman@gmail.com and ander.biguri@gmail.com

December 2022

Abstract. Krylov subspace methods are a powerful family of iterative solvers for linear systems of equations, which are commonly used for inverse problems due to their intrinsic regularization properties. Moreover, these methods are naturally suited to solve large-scale problems, as they only require matrix-vector products with the system matrix (and its adjoint) to compute approximate solutions, and they display a very fast convergence. Even if this class of methods has been widely researched and studied in the numerical linear algebra community, its use in applied medical physics and applied engineering is still very limited. e.g. in realistic large-scale Computed Tomography (CT) problems, and more specifically in Cone Beam CT (CBCT). This work attempts to breach this gap by providing a general framework for the most relevant Krylov subspace methods applied to 3D CT problems, including the most well-known Krylov solvers for non-square systems (CGLS, LSQR, LSMR), possibly in combination with Tikhonov regularization, and methods that incorporate total variation (TV) regularization. This is provided within an open source framework: the Tomographic Iterative GPU-based Reconstruction (TIGRE) toolbox, with the idea of promoting accessibility and reproducibility of the results for the algorithms presented. Finally, numerical results in synthetic and real-world 3D CT applications (medical CBCT and µ-CT datasets) are provided to showcase and compare the different Krylov subspace methods presented in the paper, as well as their suitability for different kinds of problems.

1. Introduction

Computed Tomography (CT) is a very popular imaging technique widely used in medical and scientific applications. In particular, Cone Beam CT (CBCT) has gained significant attention in the last decade, both for medicine, when low dose image guidance is required (e.g. dental imaging, image guided radiation therapy, image guided surgery), but also in scientific applications involving µ-CT, where higher doses are tolerated in favour of a better image reconstruction quality. Moreover, since many clinical applications require producing reliable images in real or near real time [1][2], there is a true need for faster available reconstruction methods. This is crucial in CBCT imaging during surgical procedures, where the long time required by most standard algorithms makes their use unfeasible in a standard clinical workflow. For example, this is the case in needle-based procedures, where fast CBCT imaging has the potential to accurately image intraoperative anatomy in close proximity to the needle [3][4] allowing for immediate adjustment in case of misplacement. Other examples can be found in image guided radiotherapy and online radiotherapy, and particularly in particle radiotherapy, where a CBCT image is taken on-site mere seconds before the radiation dose is delivered [5], with a very limited window for both reconstruction and radiation dose planning. Finally, optimization of source-detector CBCT trajectories has recently shown great promise in interventional radiology, offering a variety of benefits, including image quality improvement, FOV expansion, radiation dose reduction, metal artifact reduction, and 3D imaging under kinematic constraints. This optimization process is highly dependent on the image reconstruction speed, so the clinical implementation of such methods can only be realized with the use of fast CBCT reconstruction techniques [6][7].

In order to perform the CT reconstructions of an image from its measured x-ray projections, one needs to study and understand the properties its underlying numerical model. Mathematically, this can be formulated as finding a solution of a large-scale linear system of the form

$$Ax + e = b, \tag{1}$$

where $A \in \mathbb{R}^{N\times M}$ is the system matrix describing the measurement process, $b \in \mathbb{R}^N$ is the vector of measurements and $e \in \mathbb{R}^N$ is the modelled additive noise. Note that N is the number of detector pixels multiplied by the number of projection angles, while M is the number of voxels in the image x. For more information see, e.g., [8][9] and references therein. There are two main factors that make problem (1) very challenging to solve in practice. First, the problem is ill-posed, i.e. the matrix A has singular values that decay and cluster at zero, without an evident gap between two consecutive values. This means that the recovered solution is very sensitive with respect to small perturbations (e.g. noise) in the measurements, and therefore some regularization (replacing the original problem by a related more stable problem), needs to be applied to obtain a meaningful reconstruction. In the context of CT, the ill-posedness of the problem is related to A being a fine enough discretization of an integral operator (linear and

compact) [10], and it is accentuated when the data set is limited e.g., when only limited angle or sparse full-angle tomography measurements are available [11, Chapter 9]. The methods described in this paper provide different forms of regularization that will be explained and compared in the following sections. Second, in real-world CT applications, equation (1) can be very large-scale, so it is unfeasible to work directly with the matrix A or, in most cases, even construct it and store it.

In practical CT applications, an approximation of the solution of (1) is frequently computed using a direct method commonly known as Filtered Backprojection (FBP), or as FDK in CBCT problems, and named after Feldkamp, Davis and Kress [12]. This produces good results for mildly ill-posed problem, e.g. for high doses and independent projections. However, these algorithms can produce heavy image artifacts due to noise amplification related to the ill-posedness of the problem and mismatches between the idealistic models for the x-ray behaviour and the real measurement sampling process, see, e.g. [11]. An alternative to solve problem (1) is to use iterative methods that rely only on matrix-vector products with A and A^T to handle the large-scale nature of these matrices; hence these are also known as matrix-free methods. For mildly ill-posed problems one can expect the outcome of most used inversion algorithms to be similar [11, Chapter 9]. However, iterative methods have shown to produce reconstructions of better quality [13][14][15], particularly in the cases where there are less measurements, or they are noisier. This is especially relevant in medical applications, where reduced measurements lower the amount of damaging x-ray radiation that is given to the patient. Consequently, iterative methods are of practical relevance in clinical applications: progressively more commercial CT scanners come with iterative reconstructions due to their robust and improved image quality. Moreover, while in µ-CT the radiation dose is not harmful for the imaged object, there are cases where a sparse or low dose sampling is still required. For example, for non destructive testing of manufacturing processes, the throughput of the scanning should align with the throughput of the production, so the measuring speed limits the amount of data that can be acquired [16]. However, one needs to mention that iterative methods are slow compared to FDK: this is because FDK is computed by a ramp filter and a back-projection; whereas all existing iterative method compute, at least, one forward projection and one back-projection per iteration, and therefore each iteration requires almost the same computation time than FDK. For this reason, it is critical to make fast iterative reconstruction algorithms available for CBCT.

This work focuses on Krylov subspace methods, a family of matrix-free algorithms that are very well-known and studied in the numerical linear algebra community but that have found limited use on real-world CT applications so far. This class of methods was first introduced in the 1950s [17], but it is recently getting very popular for solving inverse inverse problems [18][19]. Conjugate Gradient Least Squares (CGLS) is the most commonly used Krylov method in applied x-ray CBCT, see, for example [20][21][22]. Moreover, it is sometimes also found in combination with Tikhonov regularization, or

within more complex minimization schemes tackling different variational regularizers [23]. Mathematically equivalent to CGLS, the more stable algorithm LSQR, has also been used for CBCT [24][25], but is by far less popular than CGLS. Other minimal residual Krylov solvers, such as modifications of the generalized minimal residual algorithm (GMRES), have been seldom used in CT [26]. In particular, we want to include in our comparisons recent developments building up from this algorithm, namely ABBA-GMRES [27][28], which support unmatched backprojectors: a very common problem in large-scale CT. In this same work LSMR is also used as a comparison. Recent developments in hybrid Krylov methods incorporating Tikhonov regularization and total variation regularization have not been used in real-world CT applications to the best of our knowledge. In terms of available (open source) software, some implementations of the described algorithms can be found along with the papers where they were presented, e.g. [27], or in the IR-tools toolbox [29], which provides many algorithm implementations for large 2D problems. However, these implementations are not suitable for large-scale (512^3 or bigger) CBCT problems. For this particular application, some Krylov methods have been implemented previously: the TIGRE [30], CIL [31] and ASTRA [32] toolboxes provide CGLS implementations, and the authors in [33] provide an implementation of CGLS and LSQR with limitations on image size when considering µ-CT scales.

The technical novelty of this paper is two-fold: 1) Applying state-of-the art Krylov subspace methods in real CT applications, some of them for the first time, 2) Providing the relevant codes within an open source framework: the Tomographic Iterative GPU-based Reconstruction (TIGRE) toolbox [30], that can be seamlessly used in any GPU supported device for arbitrarily large images as long as they can be stored and processed in the available machines. Moreover, reproducible numerical experiments are provided in synthetic and real-world 3D CT applications (for medical CBCT and µ-CT datasets) that showcase and compare the different methods presented in the paper.

In the following sections the most relevant Krylov subspace methods for 3D CT problems are described including the most well-known Krylov solvers for non-square systems (CGLS, LSQR, LSMR), possibly in combination with Tikhonov regularization, and recently developed methods that incorporate Total Variation (TV) regularization. Note that Section 2 (Methods) recalls the mathematical framework for the methods described in the paper. Reading this section is not necessary to use the algorithms as provided in the toolbox, so the authors suggest to anyone interested only in the direct applications of such methods to skip this section. Examples of results under different CT acquisition modes and samples are given in Section 3 (Results), some general guidance on the use of the different algorithms is provided in Section 4 (Discussion) and a list of the algorithms in the toolbox is provided in Section 5 (Conclusions).

2. Methods

Krylov methods are projection methods, i.e. iterative methods that, at each iteration k, are defined to find the best solution x_k belonging to a Krylov subspace of

increasing dimension according to different optimality criteria that define each particular Krylov solver. In particular, Krylov subspaces are generated by the linear combination of the first k-1 powers of a matrix acting on a vector. In this paper, we mainly focus on subspaces where A^TA acts on A^Tb, denoted as

$$\mathcal{K}_k(A^TA, A^Tb) = \text{span}\{A^Tb, (A^TA)A^Tb, ..., (A^TA)^{k-1}A^Tb\}, \tag{2}$$

or in variations thereof. Unless explicitly stated otherwise, the computational cost of the presented Krylov methods is dominated by a matrix vector product by A and A^T per iteration.

2.1. Least squares problems

Note first that (1) might not be consistent, i.e. there might not exist a solution x^* such that $Ax^* = b$, mainly due to the presence of the noise e, but also due to small differences between the discretized model and the true underlying physical model governing the measurement process. Therefore, we consider instead the following least squares problem

$$\hat{x} = \arg\min_x \|Ax - b\|_2. \tag{3}$$

Note that solving (3) corresponds to finding the best linear unbiased estimator for the solution assuming uncorrelated noise with equal variance and zero mean by the Gauss–Markov theorem, see e.g. [34]. This is frequently taken as a reasonable approximation to the noise due to its computational convenience, see e.g. [11, Chapter 2.3.2], when the noise is not 'too big', e.g. far from the low photon count limit. This is the approach that will be used in the following of this paper. It is worth mentioning that, alternatively, a more accurate quadratic approximation of the noise can be computed starting from a Poisson distribution modeling the errors for the photon counts and using a second-order Taylor expansion, see [10, Example 12.6]. This results in the variance not being the same across pixels, and can be tackled with a weighted norm in (3) (or equivalently left preconditioning in (1)). The Krylov methods in this paper can be easily adapted to this case see, e.g. [35].

When A is ill-posed, problem (3) is very sensitive to small perturbation in the measurements, so the solution of (3) might still be a bad reconstruction of the original image. Krylov methods have inherent regularization properties when combined with early stopping, displaying a phenomena called semiconvergence, i.e. the relative error norm of the solution decreases on the first iterations but starts increasing again after the optimal stopping iteration, see e.g. [9, Section 6.3]. In the following subsections the most applicable Krylov methods to solve problem (3) in the context of CT reconstruction are described.

2.1.1. CGLS Conjugate gradient least squares (CGLS), is the most used Krylov method in CT, and it dates back to [17]. It consists on applying the conjugate gradient

method to the normal equations associated to (3): $A^TAx = A^Tb$. At each iteration k, the solution of

$$x_k = \underset{x \in \mathcal{K}_k(A^TA, A^Tb)}{\arg\min} \|Ax - b\|_2 \tag{4}$$

is computed, such that the residual norm $\|r_k\|$, where $r_k = b - Ax_k$, decreases monotonically.

2.1.2. LSQR The LSQR method is based on the construction of a Krylov subspace using the Golub–Kahan (GK) bidiagonalization process [36]. This process results on a partial decomposition of A of the form $AV_k = U_{k+1}H_k$, where $H_k \in \mathbb{R}^{k+1\times k}$ is bidiagonal, and such that the orthonormal columns of V_k span the Krylov subspace $\mathcal{K}_k(A^TA, A^Tb)$, U_{k+1} has orthogonal columns and $U_{k+1}e_1 = \|b\|e_1$. Then, problem (4) can be reformulated as

$$x_k = V_k y_k \quad \text{where} \quad y_k = \underset{y\in\mathbb{R}^k}{\arg\min} \|b - AV_k y\|_2 = \underset{y\in\mathbb{R}^k}{\arg\min} \|\|b\|e_1 - H_k y\|_2, \tag{5}$$

where e_1 is the canonical vector of appropriate dimension. Even if this method has been less used in applied CT papers compared to CGLS, these two methods are mathematically equivalent, and LSQR was originally designed to provide a more stable algorithm. A detailed implementation of this method based on short recursions can be found in the original paper [36].

2.1.3. LSMR Similarly to LSQR, LSMR is also based on the construction of a Krylov subspace using the Golub–Kahan (GK) bidiagonalization process [37]. However, at each iteration, LSMR seeks a solution $x_k \in \mathcal{K}_k(A^TA, A^Tb)$ such that $\|A^Tr_k\|$ is minimized, i.e.

$$x_k = V_k y_k \quad \text{where } y_k = \underset{y\in\mathbb{R}^k}{\arg\min} \|A^Tr_k\|_2 = \underset{y\in\mathbb{R}^k}{\arg\min} \|\|A^Tb\|e_1 - \bar{H}_k^T H_k y\|_2, \tag{6}$$

where $\bar{H}_k \in \mathbb{R}^{k\times k}$ corresponds to the first k rows of the matrix H_k. Although both LSQR and LSMR converge in exact arithmetic to the same solution, see e.g. [37], they produce slightly different solutions at each iteration. Moreover, LSMR is mathematically equivalent to GMRES [38] applied to the normal equations $A^TAx = A^Tb$, and since the system matrix for the normal equations is symmetric, this is also equivalent to using MINRES [39].

2.1.4. AB-GMRES and BA-GMRES Due to how efficient implementations of the CBCT problems are coded for GPUs, the matrix B that represents the backprojection operator, i.e. the adjoint of A, is usually just an approximation of A^T [40]. This mismatch can cause the standard Simultaneous Iterative Reconstruction Technique (SIRT) family of iterative solvers to diverge, unless specific perturbations are added to stabilize the convergence, see [27]. Alternatively, the approximated transpose matrix B can be used as a right (resp. left) preconditioner for GMRES when solving problem (3), giving rise to AB-GMRES (resp. BA-GMRES) [27].

2.2. Tikhonov regularization

Another form of regularization is Tikhonov regularization, and it is perhaps the simplest and most well-known variational regularization method. It consists on computing the solution

$$\hat{x} = \arg\min_x \left\{ \|Ax - b\|_2^2 + \lambda^2 \|x\|_2^2 \right\}, \tag{7}$$

where the regularization parameter λ balances the effect of the fit-to-data term $\|Ax-b\|_2^2$ (promoting consistency of the solution with the measurements) and the regularization term $\|x\|_2^2$ (promoting regularity of the solution). If λ is chosen adequately, the semiconvergence behaviour can most times be alleviated, and the algorithms are less sensitive to early stopping; moreover, this allows for a bigger Krylov space to be built, sometimes leading to solutions of improved quality with respect to their non-Tikhonov-regularized counterparts. When λ is known, or fixed ahead of the iterations, one can apply any iterative solver (e.g. CGLS, LSQR or LSMR) to the augmented system:

$$\hat{x} = \arg\min_x \left\| \begin{bmatrix} A \\ \lambda I \end{bmatrix} x - \begin{bmatrix} b \\ 0 \end{bmatrix} \right\|_2. \tag{8}$$

For example, an LSMR implementation for fixed λ is given in the original paper [37] and compared in this study. An alternative approach is to use hybrid methods: which consist on adding Tikhonov regularization to the projected problem (5) or (6). In the case of LSQR, this is mathematically equivalent to projecting the regularized problem (8) [9, Chapter 6]. However, this is not the case for LSMR [41]. The big advantage of hybrid methods is that they provide a framework to estimate λ on-the-fly when it is not known a-priori. Even if they display a very fast convergence, the drawback of these methods is that they come with the additional cost of having to store k additional (basis) vectors for computing the solution at iteration k. This makes them suited for small to medium problems, e.g. $x \in \mathbb{R}^{512\times512\times512}$, $b \in \mathbb{R}^{512\times512\times360}$. In some cases, this could be alleviated by storing the coefficients and recomputing all the basis vectors at the end of the iterations requiring twice as many matrix-vector products with A and A^T than their non-hybrid counterparts. We provide a version of hybrid LSQR to show the performance of these methods. For more information, a great review on hybrid methods can be found in [18].

2.2.1. hybrid LSQR Using the same Krylov subspace described for LSQR and adding regularization to the projected problem (5), leads to solving, at each iteration k:

$$x_k = V_k y_k \quad \text{where } y_k = \arg\min_y \left\{ \|\|b\| e_1 - H_k y\|_2^2 + \lambda_k^2 \|y\|_2^2 \right\}. \tag{9}$$

As already mentioned, and thanks to the shift invariance property of Krylov subspaces, for fixed λ_k problem (9) is equivalent to projecting problem (8) onto the Krylov subspace $\mathcal{K}_k(A^TA, A^Tb)$, see, originally [42], or [9] for a more detailed explanation. An interesting

feature of formulation (9) is that λ_k can be computed on-the-fly at each iteration k according to a parameter choice criterion; examples of which can be found in the following section.

2.2.2. Parameter choice criteria A good choice of the regularization parameters is crucial to obtain meaningful reconstructions when dealing with ill-posed problems. In this section we focus on choices for λ_k (but note that the total amount of iterations k can also be considered a regularization parameter for regularization by early-stopping). In the following, we provide the description of two of the most simple and commonly used regularization parameter choice criteria. This is by no means an exhaustive list of the available options and we point the interested reader to the reviews in e.g. [18][19].

If a good estimate of the norm of the error $\|e\|$ is available, a very popular and reliable parameter choice criterion is the Discrepancy Principle (DP) [43]. This method is based on the idea that

$$\|Ax^{exact} - b\|_2^2 = \|Ax^{exact} - b^{exact} - e\|_2^2 = \|e\|_2^2 = \frac{\|e\|_2^2}{\|b\|_2^2}\|b\|_2^2 = \mathrm{nl}^2\,\|b\|_2^2 \tag{10}$$

so at each iteration, λ_k is chosen so that

$$\lambda_k = \arg\min_{\lambda}\{\|Ax_\lambda - b\|_2^2 - \mathrm{nl}^2\,\|b\|_2^2\}. \tag{11}$$

Alternatively, one can use parameter choice rules that do not use any information about the noise e, also known as "heuristic methods". In particular, we provide an implementation of the Generalized Cross Validation (GCV) parameter choice criterion, which relies on cross validation: a well known statistical tool used to predict possible missing data values. In this case, each of the components of the vector of measurements b is estimated using the rest of components, and the regularization parameter λ_k associated with the best predicted values is taken at each iteration. In practice, for hybrid LSQR, using GCV involves solving the following minimization:

$$\lambda_k = \arg\min_{\lambda} \frac{\left\|(I - H_k H_{k,\lambda}^{\dagger})\,\|b\| e_1\right\|_2^2}{\mathrm{tr}((I - H_k H_{k,\lambda}^{\dagger}))^2} \quad \text{where } H_{k,\lambda}^{\dagger} = (H_k^T H_k + \lambda^2 I)^{-1} H_k^T. \tag{12}$$

Note that this can be generalized to other Krylov methods (e.g. LSMR) by replacing the projected matrix and right hand side by corresponding ones (see, e.g. [41]).

2.3. Total Variation (TV) regularization

Total variation is a very common variational regularization scheme that promotes piecewise-constant reconstructions by favouring solutions with a sparse gradient. This is very popular in imaging problems as it contributes to preserve edges in the reconstructed image. In this paper the discrete isotropic total variation in 3D is considered:

$$TV(x) = \sum_i \sqrt{[D_l x]_i^2 + [D_j x]_i^2 + [D_k x]_i^2} = \left\| \sqrt{[D_l x]_i^2 + [D_j x]_i^2 + [D_k x]_i^2} \right\|_1, \tag{13}$$

where D_l, D_j, D_k refer to the finite difference approximations of the three directional derivatives for the 3D image x. A popular approach to solve the TV problem using Krylov methods is to re-write the TV regularization term using a weighted 2-norm:

$$\hat{x} = \arg\min_x \left\{ \|Ax - b\|_2^2 + \lambda^2 TV(x) \right\} = \arg\min_x \left\{ \|Ax - b\|_2^2 + \lambda^2 \|W(Dx)Dx\|_2^2 \right\}, \tag{14}$$

where D is the 3D discrete derivative operator and $W(Dx)$ is a (diagonal) weighting matrix that depends on Dx. Then, the functional in (14) can be approximated locally by a sequence of quadratic functionals, giving rise to a sequence of problems of the form:

$$x^{(k)} = \arg\min_x \left\{ \|Ax - b\|_2^2 + \lambda^2 \|L^{(k)} Dx\|_2^2 \right\}, \tag{15}$$

where $L^{(k)}$ are approximations of $W(Dx)$ of improving quality. This scheme is called iteratively reweighted norm (IRN) and was first used in combination with TV in [44] for 2D imaging problems. In the following, two algorithms that (partially) solve the problems in (15) to approximate TV regularization are described.

2.3.1. CGSL-TV The sequence of problems (15) can be solved in an inner-outer scheme fashion where, at each outer iteration, the computed solution $x^{(k)}$ is used to update the weights $L^{(k+1)} = W(Dx^{(k)})$. Following [44], an adaptation of this method for 3D using CGLS in the inner iterations, is provided in this paper. This scheme has provable convergence guarantees, but requires λ to be known a-priori and can be very computationally expensive due to its inner-outer scheme nature. Other variations of this method have been implemented using other Krylov methods for the inner iterations, e.g., in combination with LSQR [45].

2.3.2. hybrid fLSQR An equivalent formulation to (15), dropping the (k) upper-script to ease the notation so that $L = L^{(k)}$, is to solve

$$\hat{x} = L_A^\dagger \bar{y}_L + x_0, \quad \text{where} \quad \bar{y}_L = \arg\min_{\bar{y}} \left\{ \|AL_A^\dagger \bar{y} - \bar{b}\|_2^2 + \lambda^2 \|\bar{y}\|_2^2 \right\}, \tag{16}$$

where $L_A^\dagger$ is the A-weighted pseudoinverse of L, defined as $L_A^\dagger = [I - (A(I - L^\dagger L))^\dagger A] L^\dagger$ ($L^\dagger$ denotes the Moore-Penrose pseudoinverse of L); x_0 is the component of the solution $\hat{x}$ in the null space of $L^{(k)}$ and $\bar{b} = b - Ax_0$. The matrix $L_A^\dagger$ can now be considered as an (iteration dependent) right preconditioner and incorporated into the space of the solutions using flexible Krylov methods, see, e.g. [46][47][48]. This strategy circumvents the need for an inner-outer scheme, and provides a much faster convergence than TV - CGSL. Moreover, in a hybrid fashion, it allows for the regularization parameter $\lambda = \lambda_k$ to be computed on-the-fly throughout the iterations. However, flexible Krylov methods require storing all the computed basis vectors so that the memory requirements increase with the number of iterations. The algorithm provided in this paper is an adaptation from [48], using different boundary conditions for the discrete derivative operator approximation and extending it to 3D.

3. Numerical experiments

In this section, three representative numerical experiments are presented and discussed to illustrate different aspects of the Krylov subspace methods described in this paper. The aim of this section is to provide greater depth in the understanding of the behaviour of the described algorithms in practice, which can be used as a blueprint for 'what to expect' when using them in other data-sets, rather than producing and exhaustive evaluation of Krylov methods against the entire reconstruction literature. Due to the large number of reconstruction algorithms, the difficulty of obtaining real datasets, and the task-specific nature of the quality metrics, this is beyond the scope of this paper.

The first presented example consists of synthetic CT data, where the true image is known and the results can be analyzed and discussed in detail: providing a comprehensive comparison including relative error and residual norm histories. The second experiment has the aim of showing the typical performance of these methods on real data: it consists of a scan of the Alderson head phantom obtained in a Philips Allura medical CT scanner that is reconstructed with full sampling and under-sampled projections. Finally, a bumblebee image obtained with an industrial Nikon CT scanner is reconstructed for some of the algorithms, in a real large-scale problem. Note that the large-scale nature of the CBCT image reconstruction problems means that the implementations are often in single precision floating point arithmetic.

The first two experiments were carried out in a laptop with a Intel Core i7-7700HQ with 16GB of RAM and a GTX 1070 NVIDIA GPU. The µ-CT reconstruction was performed in a machine with an AMD EPYC 7352 with 126GB of RAM and 4 NVIDIA Quadro RTX 6000.

3.1. Comprehensive convergence comparison on synthetic data

In this experiment we explore the behaviour of the algorithms presented in this paper in the context of 3D CBCT, using the available implementation in the TIGRE toolbox. Since this is a simulated toy example, both the relative residual norms, i.e. $\|Ax_i - b\|/\|b\|$, and relative error norms $\|x_i - x_{gt}\|/\|x_{gt}\|$ (for a given iteration number i and a ground truth image x_{gt}) can be computed. Relative residual norms are a natural metric to understand the behaviour of the presented algorithms, as they are the values, at each iteration, of the model fit, rescaled to ease comparison. For methods without further regularization, this coincides with the value of the objective functional (3) that we are minimizing, so this is a good indicator of the convergence of the algorithm to the minimizer of the objective functional. However, for ill-posed problems, this might be a bad indicator of the problem converging to a good approximation of the solution of the original problem (1). The relative error norm is used in this case as a standard application-agnostic metric to understand how close the reconstruction is to the true solution, along with a qualitative inspection of the results.

The data presented in this example concerns the measurements of a synthetic

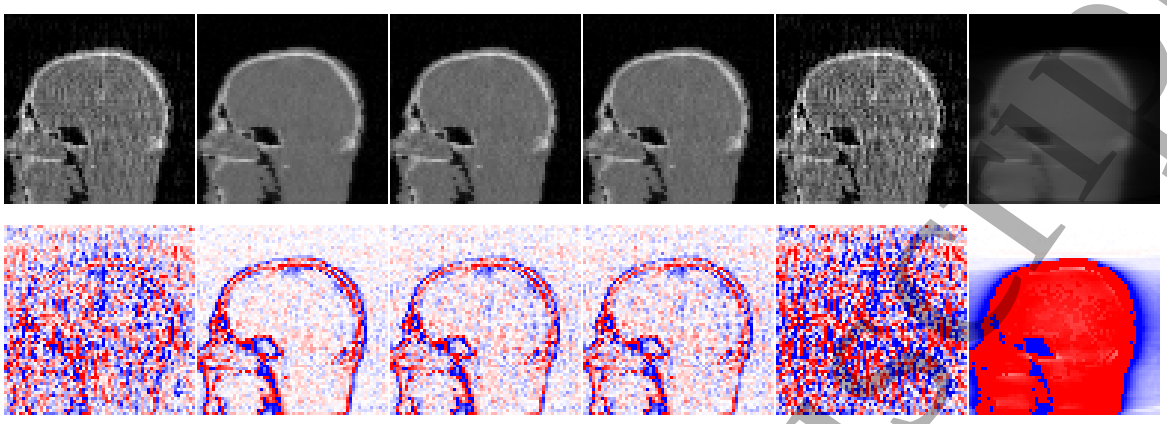


Figure 1: Reconstruction of phantom head data using different LSQR versions and parameter selection procedures. (top row) slice of final images, shown in range [0, 1] mm^{-1} (bottom row) difference images w.r.t. the ground truth, shown in range [-0.1, 0.1] mm^{-1}.

dataset of a human head of size 64×64×64 with a detector of size 128×128 pixels, using TIGRE's default geometry mimicking a medical CT scanner‡. The results for this experiment are presented an analysed in three different subsections. First, the results for LSQR and hybrid LSQR (using different regularization parameter choices) are displayed to illustrate the typical behaviour of Krylov methods for CT problems, and can be observed in Figures 1 and 2. Second, the Krylov methods presented in this paper for the least squares problems with and without Tikhonov regularization are shown in Figures 3 and 4. Finally, a sparse-view version of the same simulation is used to showcase the Krylov methods that enforce TV regularization.

3.1.1. Illustration of Krylov methods' typical behaviour In this experiment, 60 equidistant angles spanning the full circular range are simulated, with added Poisson noise (assuming an air photon count of $I_0 = 1 \times 10^5$ in each pixel) and Gaussian noise (with standard deviation of $\sigma = 0.5$) which model both the photon and electronic noise expected in a CT scanner [49][50]. The reconstructions can be observed in Figure 1, while relative residual and error norms are displayed in Figure 2.

First, we can observe that LSQR undergoes semiconvergence. As briefly mentioned in Section 2.1, this is characterized by the decrease of the relative error norm in the first iterations until it reaches a minima after which it starts increasing (see Figure 2(c)); while the residual norm is in exact arithmetic minimized at each iteration and therefore decreases monotonically throughout the iterations (see Figure 2(a)). This phenomenon is the reason why early stopping is crucial to obtain a good reconstruction in inverse

‡ Source detector distance of 1536 mm, source axis of rotation distance of 1000 mm, detector size of 409.6 × 409.6 mm and an image spanning 256 × 256 × 256 mm.

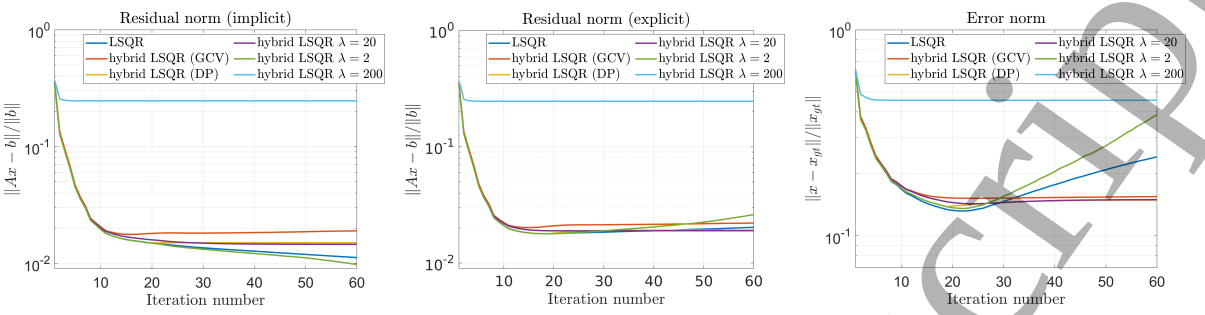


Figure 2: (a) Implicit relative residual norms, (b) computed relative residual norms and (c) relative error norms for the algorithms of interest, per iteration.

problems with noisy measurements when using an iterative solver that acts directly on the least squares problem (3).

Second, we want to illustrate how the choice of a good regularization parameter is crucial to obtain a meaningful reconstruction when solving the Tikhonov problem (7). As can be observed in both the reconstructions (Figure 1) and the relative error norm histories (Figure 2(c)), the semi-automatic parameter choice criteria provided in this implementation find appropriate parameters λ_k at each iteration to obtain a good reconstruction without fine tuning. Alternatively, one can choose a parameter λ ahead of the iterations. In this case, note that an under-regularized problem (see Figure 1 for $\lambda = 2$) will produce a noisy-looking image (in which case one should re-run the algorithms using a higher value for the parameter λ); while an over-regularized problem (see Figure 1 for $\lambda = 200$), will produce an overly smooth reconstruction (in which case one should use a smaller value for λ).

Last, a very interesting thing to note is the mismatch between the theoretical or implicit residuals in Figure 2(a), i.e. computed using $\|\|b\|e_1 - H_k y_i\|/\|b\|$ or mathematical recurrences (see [36] for LSQR), and the residuals computed explicitly throughout the iterations in Figure 2(b), i.e. using $\|Ax_i - b\|/\|b\|$ directly. This can be due to loss of orthogonality (mainly attributed to the mismatched backprojector) or to an accumulation of numerical errors and precision loss (most objects are stored in single precision floating point arithmetic, with large differences in the order of magnitude of the parameters). Note that this happens both for the algorithms that incorporate re-orthogonalization and for the ones that do not. As this mismatch flags a deviation of the algorithm from its expected behaviour in exact arithmetic, TIGRE explicitly computes the residual norms at each iteration and stops the algorithm once they increase.

3.1.2. (hybrid) Krylov methods for least squares problems This experiment concerns a dataset with 180 equidistant angular projections with the same noise distribution used in the previous section ($I_0 = 1 \times 10^5$, $\sigma = 0.5$). The results for all the algorithms presented in this work for the least squares problem with or without Tikhonov regularization,

are shown for a maximum of 60 iterations. As a baseline, the results are compared to the solutions computed with SIRT: a particular choice from the most commonly used family of algorithms in CT, the SIRT-like family (SIRT, OS-SART, SART, etc) [8], which is computationally equivalent to the Krylov methods used in this example (i.e., they have an equivalent amount of flops per iteration). Figure 3 shows a slice of the reconstructed image obtained using the different methods on top of its corresponding error (difference between the reconstructed slice and the ground truth). Figure 4 shows the relative residual norm and relative error norm histories for all the algorithms against the number of iterations.

In Figure 4 one can observe the very fast convergence of Krylov methods: both in terms of the relative residual norm and of the relative error norm. In this particular experiment, between 10 and 20 iterations of the compared Krylov subspace methods are sufficient to obtain a good reconstruction of the original image while, after 60 iterations, SIRT has still not converged and has failed to compute meaningful reconstruction. It can also be observed that the different Krylov methods perform similarly, with LSQR producing results of slightly better quality than CGLS in terms of error norm.

For this particular example, the iterations are stopped early if the norm of the explicit residual increases between two consecutive iterations, as this is a sign of loss of orthogonality in the basis vectors or of the accumulation of computational errors. This happens for most compared Krylov subspace algorithms, and note that for CGLS, LSMR and LSQR, as a side-effect, this leads to regularization by early stopping and avoids the semiconverge behaviour. However, for these algorithms, one should stop the iterations early even when the explicit residual norm does not increase (for example, by monitoring the stabilization of the residual norm or using other stopping criteria). It is also remarkable to observe that the AB/BA-GMRES algorithms less commonly display increasing residuals, as they alleviate the problems associated with mismatched backprojectors.

3.1.3. Krylov methods for total variation regularization For this experiment, the simulated CT measurements of the dataset described in the previous section are reduced and correspond to 60 equidistant projections: generating a more ill-posed problem [11, Chapter 9]. Moreover, the Poisson noise for this problem is increased so that $I_0 = 1 \times 10^4$ in each pixel. In this case, more prior information on the solution is needed to obtain a good reconstruction of the original image, and therefore the described methods involving TV regularization become more meaningful. In particular, CGLS, CLGS with TV regularization and hybrid fLSQR with TV regularization are showcased in this experiment. The reconstructions obtained by these methods after 60 iterations can be seen in Figure 5, where the smoothing but edge preserving behaviour of the TV regularization is visibly clear.

Figure 6 shows the relative residual norms and the relative error norms throughout the iterations for the compared methods. In this example it can be clearly observed that CGLS semiconverges due to the ill-posedness of the problem and the noise in the

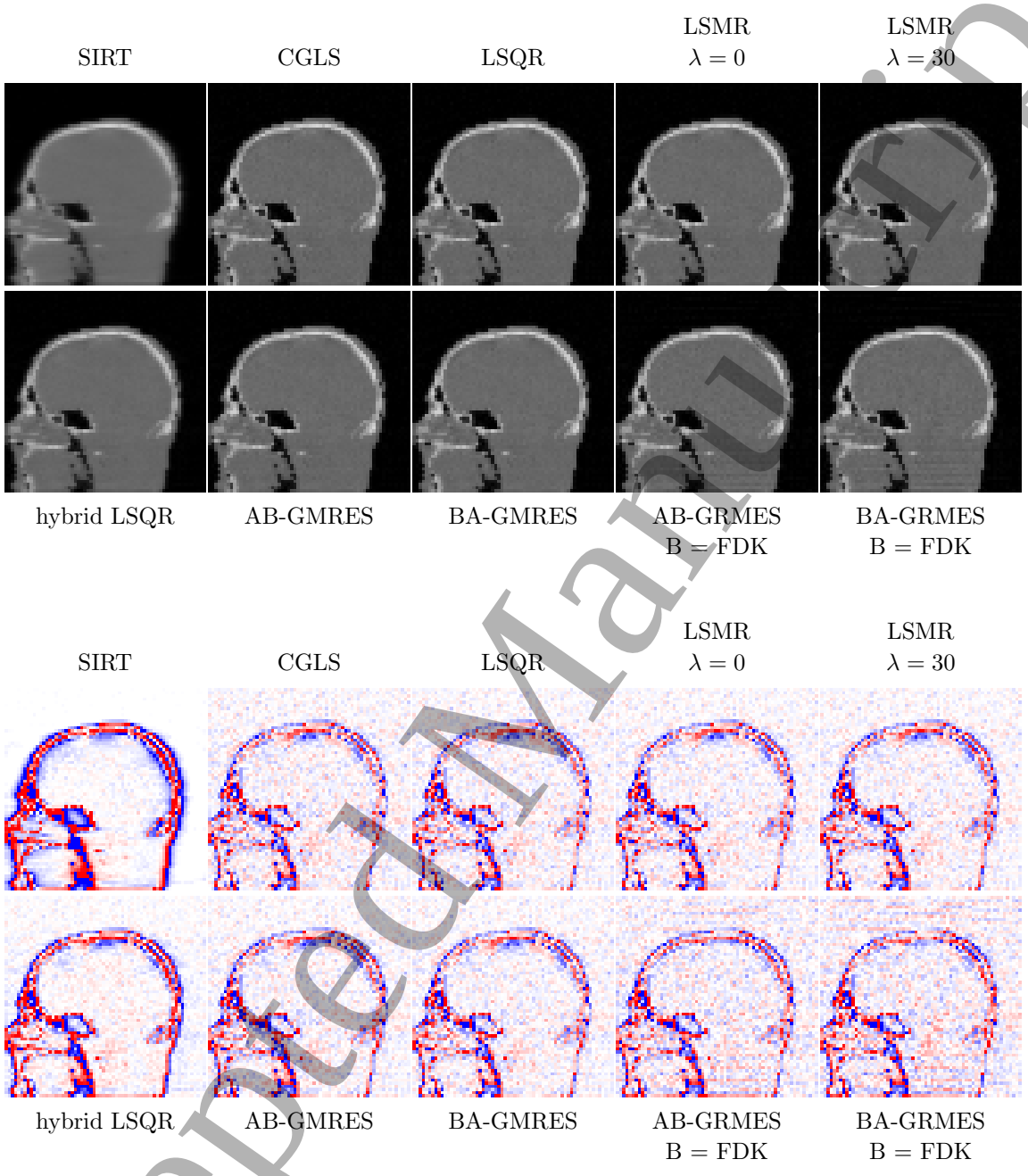
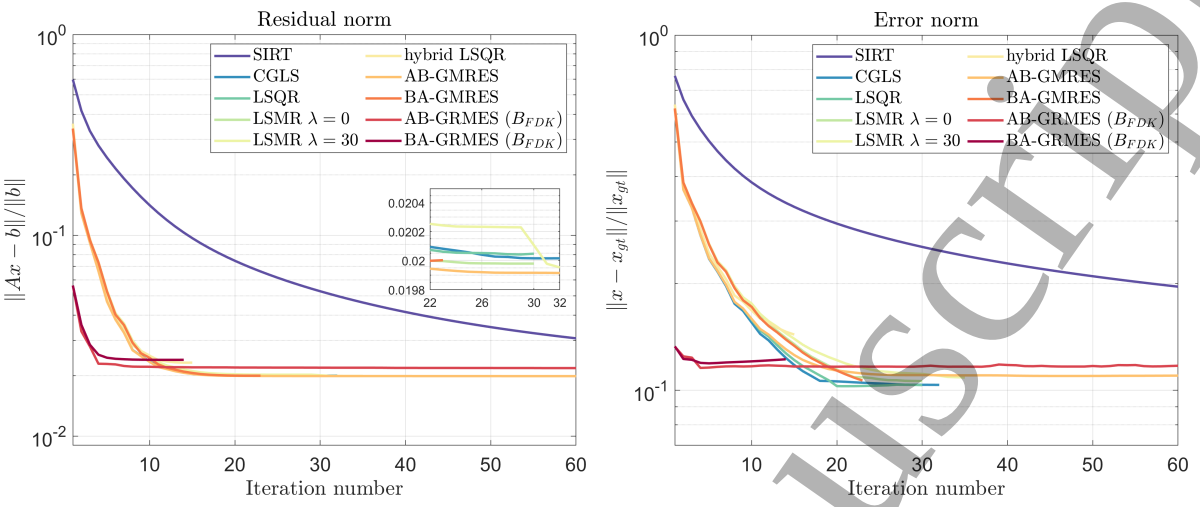


Figure 3: Reconstruction of phantom head data using several Krylov methods (top) slice of final images, shown in range [0, 1] mm^{-1} (bottom) difference images w.r.t. the ground truth, shown in range [-0.1, 0.1] mm^{-1}.



(a) (b)

Figure 4: (a) Relative residual norms and (b) relative error norms for the compared algorithms, per iteration.

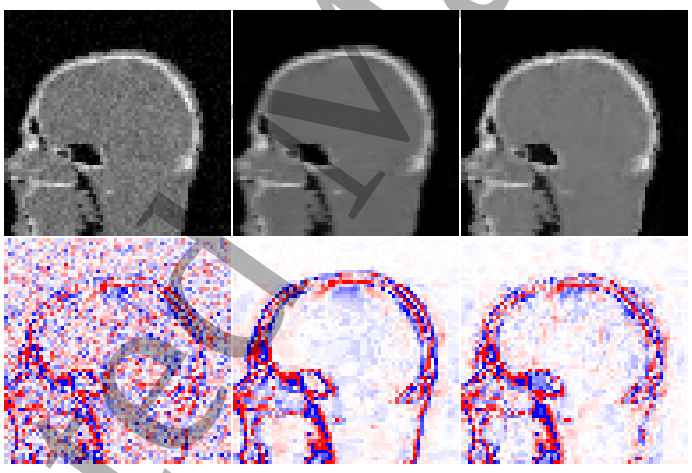


Figure 5: Reconstruction of phantom head data using CGLS and TV regularized Krylov methods. (top row) slice of final images, shown in range [0, 1] mm^{-1} (bottom row) difference images w.r.t. the ground truth, shown in range [-0.1, 0.1] mm^{-1}.

data. It is also important to explain that the behaviour of CGLS-TV in terms or relative residual and error norms is expected. Here, the 'peaks' correspond to the starts of each new cycle of inner iterations, also known as cold restarts. For this particular experiment, the number of inner iterations in each cycle is chosen a-priori to be 12 iterations, but this could also be set adaptively using a stopping criterion for the inner iterations. As long as the number of inner iterations is sufficiently large, this algorithm produces very good reconstructions with the properties expected of TV regularized solutions (this is especially desirable for highly noisy datasets). The experiment shows that likely three

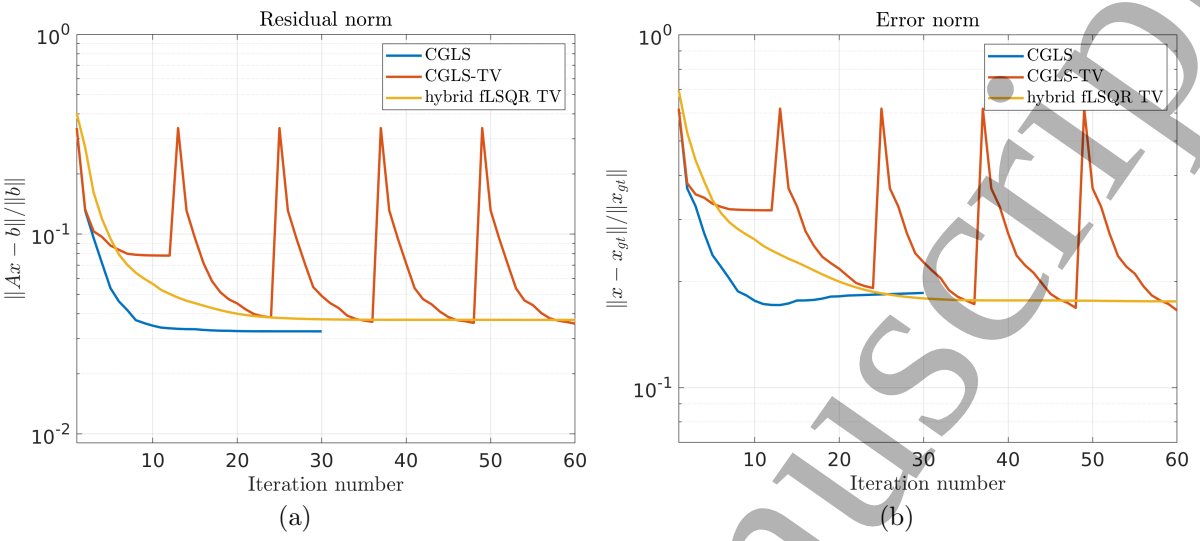


Figure 6: (a) Relative residual norms and (b) relative error norms for the compared algorithms, per iteration.

outer iterations (36 iterations in total) would be sufficient for this particular example. Finally, even if hybrid fLSQR-TV displays a slower decay of the residual norm, it still produces fastly decreasing error norms compared to CGLS-TV, and does not exhibit semiconvergence. Note that this method does not require to set a number of inner iterations, and the regularization parameter can be chosen semi-automatically on-the-fly. However, it requires storing all the generated basis vectors and has the additional cost of an (approximated) matrix-vector product with $L_A^\dagger (L_A^\dagger)^T$ at each iteration (in the codes provided, this is done efficiently using an iterative method).

3.2. Medical CT experiments

This experiment concerns a medical imaging application and has the aim to highlight the performance of Krylov subspace methods on real data. In particular, the computational times for the different algorithms are given in this example to highlight the fast convergence of Krylov methods. Since there is no 'ground-truth' for this experiment we assess the reconstructions based on a qualitative inspection of the results compared to FDK, and on the evaluation of the relative residual norm history (i.e. the relative residual norm stabilizing close to convergence).

The dataset used in this experiments consists of the Alderson head phantom measurements, acquired on a Philips Allura FD20 Xper C-arm with source settings of 80 kV and an exposure of 350 mAs, spanning a 210° angular range. Projections of size 512×512 have been used to reconstruct an image of size 256×256×200 voxels. Figure 7 shows two views of the image reconstructions given by different algorithms. The following explanations and comparisons are applied to both the sagittal plane (Figure 7(a)) and the transversal plane (Figure 7(b)) of the different reconstructions. The first

row shows images reconstructed by FDK (considered clinical standard) and SIRT with 30 and 150 iterations, respectively. Note that the choice of 30 iterations for SIRT is taken to match the computational time required for Krylov methods to obtain a meaningful reconstruction (it can be observed that the quality of the reconstruction using SIRT in this case is not very good), while the choice of 150 iterations is taken so that SIRT displays an equivalent quality of the reconstructions than Krylov methods (taking 3 minutes and 50 seconds, almost 8 times slower than Krylov methods). In the first column of the second row, OS-SART (the ordered subset version of SIRT), is shown after 60 iterations (chosen to reach a reasonable convergence). Albeit the number of required iterations is smaller than for SIRT, OS-SART takes 6 min 30 seconds to reconstruct this image§. In the second and third columns of the second row, the reconstructions obtained using CGLS and LSQR are shown after 30 iterations (corresponding to 30 seconds of run-time). The third row, from left to right, shows the reconstructions obtained using LSMR (with $\lambda = 0$), LSMR (with $\lambda = 30$) and hybrid LSQR, all of them after 30 iterations and 30 seconds of run-time. The reconstructions obtained with the studied iterative methods look less grainy than the baseline reconstruction obtained using FDK.

In this experiment one can observe that iterative methods produce image reconstructions of similar or higher quality than the clinical standard FDK. Moreover, Krylov subspace methods are able to do so in significantly less computational time compared to other classic iterative reconstruction methods. This is of particular use in clinical CT, where lower reconstruction time is needed to maximize throughput (i.e., the number of images and actions over them that can be processed per unit of time).

In the following, the reconstruction results for the same example with a fifth of the projection data are shown to simulate a sparse sampling CT scan. Figure 8 displays the results in the same order and for the same number of iterations already described for Figure 7. In this scenario, the Krylov subspace methods produce a good reconstruction in less than 15 seconds. Note that using SIRT in a comparable time (30 iterations) produces overly smooth reconstructions, i.e. they appear less noisy but the lack of sharpness in the edges might lead to the loss of important features in the image.

Finally, TV regularized Krylov algorithms are used in this experiment with under-sampled projections to showcase the impact of this type of regularization on challenging CT scans. Figure 9 shows, for comparison, the reconstructions obtained using LSQR after 30 iterations (same as in Figure 8), OS-ASD-POCS, an ordered subset version of a well known TV regularized algorithm in tomography [51] after 60 iterations; and CGLS with TV regularization (2 outer and 15 inner iterations). For this experiment, the running time for CGLS-TV is 1 minute, while the running time for OS-ASD-POCS is 2 minutes. It is not straightforward to establish a fair comparison purely between these two algorithms in terms of reconstruction quality, as they require the choice of

§ This is specific for the particular TIGRE implementation. Faster subset algorithms can be developed using specific implementations that minimize CPU⇔GPU memory transfers. However, they require a larger amount of computations per iteration than other iterative methods, so they will still be slower than the other algorithms shown in this paper.

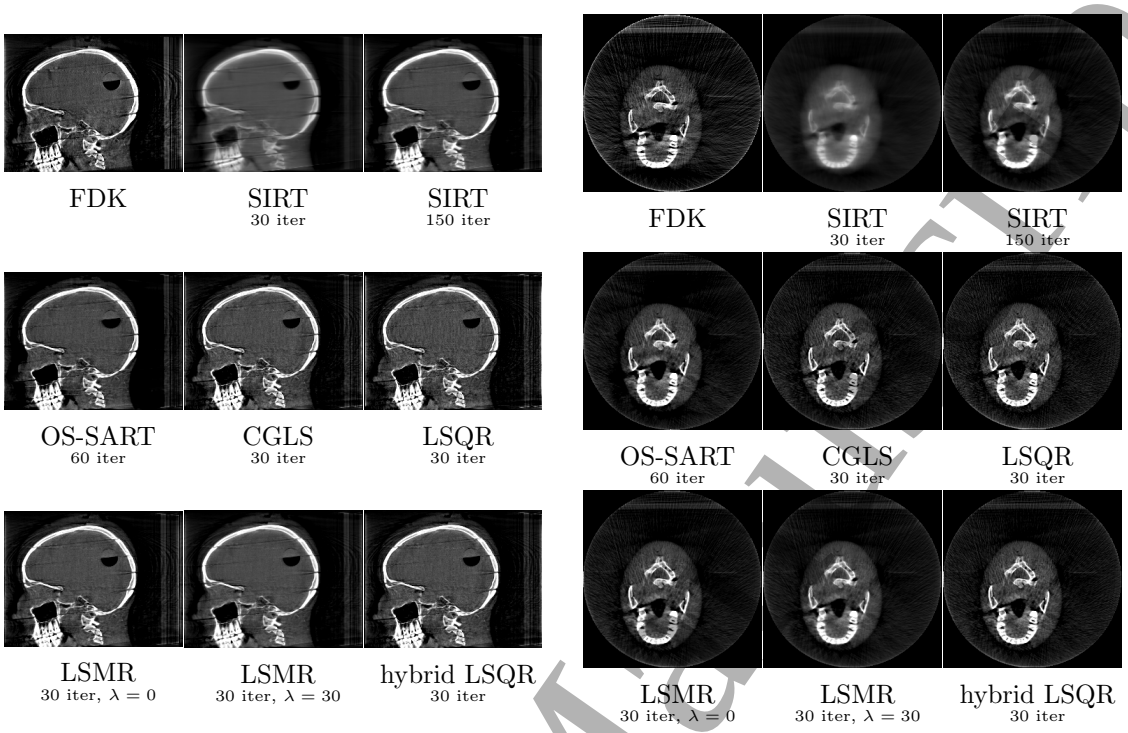


Figure 7: Reconstruction of the Alderson head phantom acquired on a Phillips Allura FD20 Xper C-arm, using 289 projections. Image shown in range [0, 3] mm^{-1} for the (left) sagittal plane, (right) transversal plane. SIRT 30 iterations and the Krylov subspace algorithms terminate within 35 seconds, while SIRT 150 iteration takes 3 minutes 50 seconds and OS-SART 6 minutes 30 seconds to converge to a solution of comparable quality.

different regularization hyperparameters (and there is no direct translation between the parameters for both of these algorithms). These will have a great influence in the reconstruction, balancing a closer reproduction of the fine detail features and a general smoother piece-wise constant appearance of the images. However, the results show that CGLS-TV produces good results (relatively smooth with sharp edges), while preserving the finer details of the image structures. Note that the hybrid fLSQR algorithm was not used in this experiment because the high memory needs of this algorithm were too big for the machine in which the experiments were run: this algorithm, in its current state, may not be suitable for a medical-size dataset.

3.3. μ-CT scan

This experiment showcases the use of the methods presented in this work in very large-scale problems where the radiation dose is not an issue. The scanned object is a wild buff-tailed bumblebee (*bombus terrestris*)‖, scanned on a Nikon HMX 225 kVp CT

‖ The unfortunate individual was found dead in the X-ray CT laboratory after getting trapped inside and became an essential part of the laboratory as an independent research dataset.

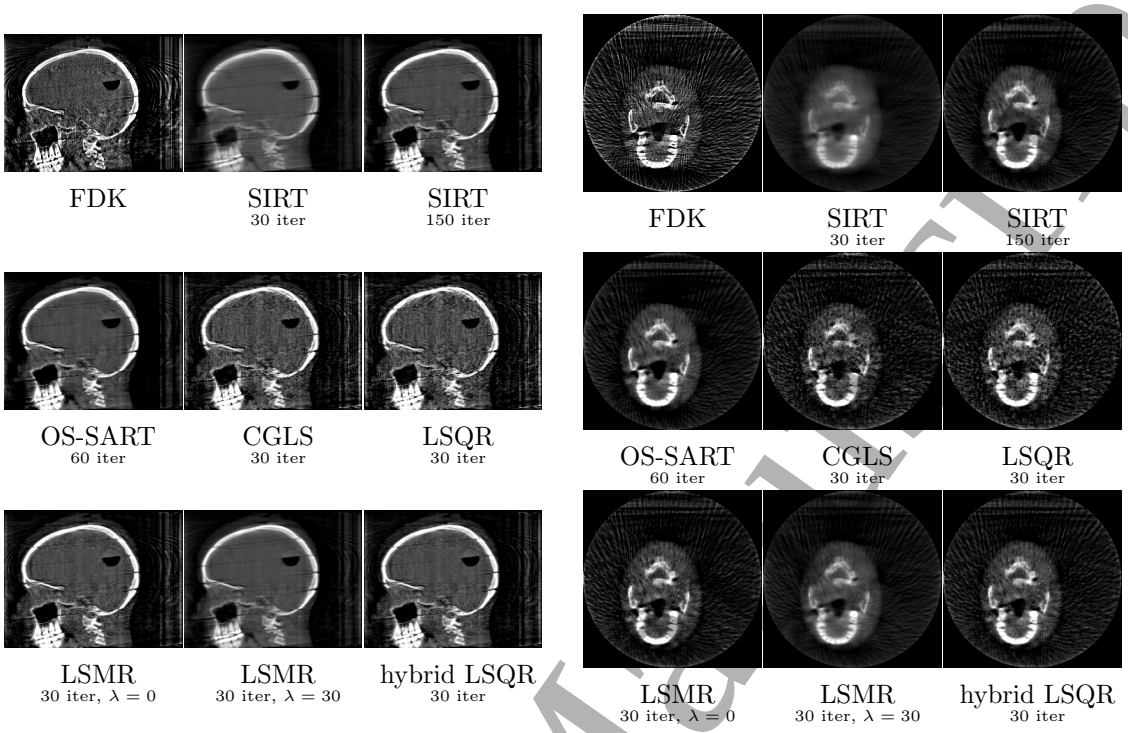


Figure 8: Reconstruction of the Alderson head phantom acquired on a Phillips Allura FD20 Xper C-arm, using 58 projections. Image shown in range [0, 3] mm^{-1} for the (left) sagittal plane, (right) transversal plane. SIRT 30 iterations and the Krylov subspace algorithms terminate within 15 seconds, while SIRT 150 iteration takes 1 minutes 30 seconds and OS-SART 1 minutes 50 seconds to converge to a solution of comparable quality.

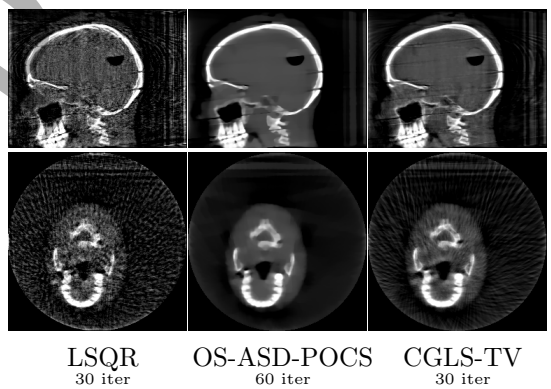


Figure 9: Reconstruction of the Alderson head phantom acquired on a Phillips Allura FD20 Xper C-arm, using 58 projections. Image shown in range [0, 3] mm^{-1} of (top row) the sagittal plane (bottom row) the transversal plane.

scanner at 40 kVp with a molybdenum target. The detector was a Perkin Elmer 1621 with a gadolinium oxysulphide scintillator. The detector is of size 2000×2000 and we used 256 projections uniformly distributed around the circle. The reconstructed image is $1400 \times 1400 \times 2000$ with a resolution of 11.8 µm per voxel. Figure 10 shows the FDK reconstruction and LSQR reconstruction with 20 iterations (15 minutes of computational time). The different nature of the reconstructed images can be seen. In particular, the attenuation values of the tissue of the bumblebee are more uniform in LSQR (uniformity in the tissues is the expected result) and some features are better distinguished from the noise, particularly noticeable in the middle thin string-like structure in the zoomed-in area. However, this reconstruction also highlights one potential issue with iterative methods when mismatches in data consistency are present in the measurements. In particular, for some acquisitions where the edges of the projections might have errors due to e.g. photon starvation, or partial views of samples, iterative algorithms might produce artifacts that propagate through the image, as one can see in the stripe artifacts arising near the head (right part) of the Bumblebee, to the point where some features are considerably worse, or missing. In particular, the acquisition process for this dataset: highly sampled but with limited field of view (not the entire sample is in the imaging domain) favours FDK reconstructions over iterative methods. Therefore, while it is important to remark that some of the artifacts produced with Krylov methods can be easily alleviated using data acquisition techniques that are tailored for iterative algorithms, it is also important to note that the acquisition process is a very important factor to consider when reconstructing already available datasets. However, we have shown that it is feasible to use Krylov subspace methods in really large-scale settings, so exploring different data acquisition techniques becomes a relevant problem.

4. Discussion

This section provides a discussion on some aspects reported in the numerical experiments, as well as some guidance on how to use some of these methods, with the aim of explaining potential issues that one might encounter when applying these algorithms to other datasets.

One of the results that is mentioned in this paper is the fact that, in practice, the real residual norm can increase throughout the iterations due to a loss of orthogonality (this is not expected in exact arithmetic for the methods that theoretically minimize the residual at each iteration) or due to an accumulation numerical errors (this behaviour is also observed for the algorithms that incorporate re-orthogonalization). As described in the previous sections, this is mostly due to the use of an unmatched backprojector. The TIGRE toolbox provides an approximation of a matched backprojector [52] that mitigates the diverging behaviour in the implicit and explicit residual norms for the Krylov methods. Similarly, this is also mitigated when using algorithms that incorporate re-orthogonalization (but these come with the added cost of having to store all the computed basis vectors). An even better solution would involve implementing matched

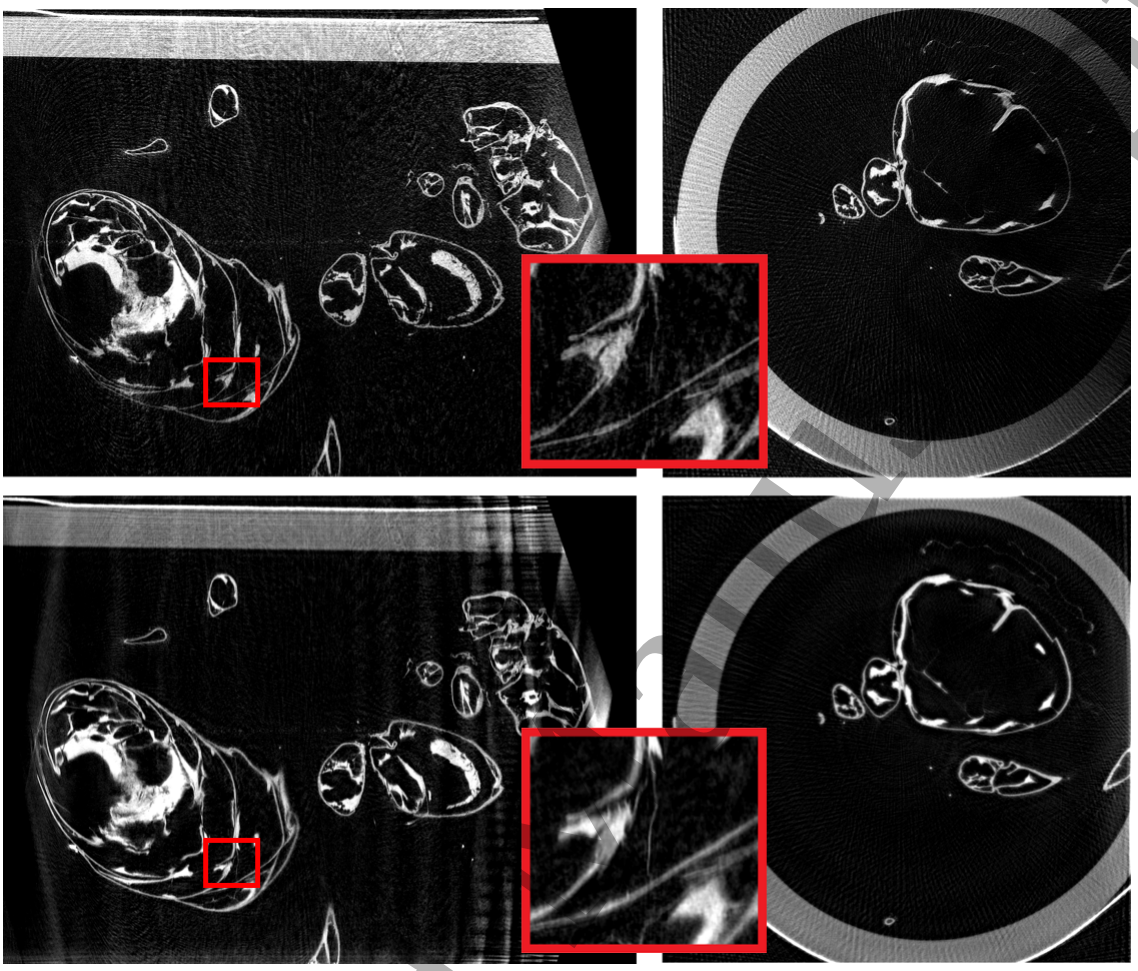


Figure 10: FDK (top) and LSQR 20 iterations (bottom) reconstruction of a µ-CT scan of a Bumblebee. The reconstruction is of size $1400 \times 1400 \times 2000$, using 256 projections of a 2000×2000 detector. Zoomed area shows the different nature of the reconstructions, highlighting the noise rejection nature of iterative reconstruction. The small line in the center of the zoomed area can be better distinguished in the LSQR image. Image displayed at [0, 0.1] mm^{-1}.

projection/backprojection operators, such as the distance driven projectors [53], or pixel driven matched projector approximations [54]. Further research on the impact of numerical precision on Krylov methods would also be beneficial.

A natural question that can arise from applied scientist is which algorithm is the "best". First, this is an unanswerable question in general, as the algorithm choice (as well as the desired type and level of regularization) should heavily depend on the specific problem in mind and the purpose of the reconstruction. For example, if one would only want to have a general understanding of the structure of the internals of the Bumblebee in Figure 10, FDK provides a sufficiently good image quality. Perhaps this would not hold if a quality segmentation of the image was required. Often the right algorithm choice for the reconstruction depends on what the image will be used for, instead of

some arbitrary image quality metric. In fact, there is a nuanced discussion to have about the use of image quality metrics (such as SSIM, UQI, RMSE) to evaluate the results, as these are based on preconceptions on what natural images should look like, rather than, for example, clinical relevance [55, 56]. Following from this, we decided to avoid using these metrics in this paper to evaluate the performance of the algorithms. Instead, we opted for qualitative metrics for the quality of the reconstructions (choosing "meaningful reconstructions" leveraging a small error norm and desirable observable properties such as smoothness of the background and sharpness of the edges) and standard quantitative metrics for the convergence of the algorithms across the iterations. As a matter of fact, the objective of this work is not to "rank" the described algorithms, but to compare their properties and to supply easy to use and reproducible tools for exploration.

Some tips can be provided on the general use of these algorithms. It is recommended to use LSQR over CGLS, as it is a more stable algorithm but they are mathematically equivalent. In general, for severely undersampled CT measurements, especially with high noise levels, explicit regularization is recommended. In particular, TV regularization can be very beneficial to promote sharp edges (note that this is not only true for Krylov methods). Moreover, high regularization will produce less grainy reconstructions, but over-regularizing might lead to loosing tissue/material texture properties; this can be beneficial in some contexts, e.g. segmentation or classification, but a problem in other contexts, e.g. when the finer details are important and the noisy appearance of the reconstruction is not a problem. Finally, if enough memory is available (an image per iteration), algorithms with explicit re-orthogonalization are recommended, such as AB/BA-GMRES, to mitigate the problems derived from the loss of orthogonality.

It is also important to state that this is a representative but by all means not comprehensive list of all Krylov methods and their corresponding features; such as stopping criteria, see, e.g. [57] or parameter selection criteria, see e.g. [18][19]. Similarly, many direct, variational, and more recently machine learning methods are being developed. The authors encourage the public to contribute to this work by submitting new algorithms or improving the ones available at the TIGRE toolbox.

5. Conclusions

This work describes and compares a variety of Krylov subspace methods for applied large-scale 3D CT and CBCT reconstruction, some of them used in this context for the first time. In particular, the methods included in this work are summarized in Table 1.

The considered Krylov methods are compared and discussed in the numerical experiments, see Section 3, where it can be clearly observed that the main strength of Krylov subspace methods is their fast convergence compared to the most commonly used SIRT-like methods in iterative CT reconstruction. This is of crucial importance in medical applications, for example in image guided therapies where almost real time

reconstructions are needed, but also in industrial applications where a high number of iterations is unfeasible due to the big dimensionality of the problems.

Finally, all the results shown in the paper are reproducible, and all the methods are provided as open source and freely accessible algorithms within the framework of the TIGRE toolbox. Some guidance on how to use these methods and a small discussion on potential results and issues one might encounter when using them on other datasets is given in the discussion, see Section 4. All the methods presented in this work can be found at `github.com/CERN/TIGRE` under a permissive BSD-3 clause license.

Acknowledgements

MSL gratefully acknowledges support from the CMIH, University of Cambridge. AB acknowledges the support of EPSRC grant EP/W004445/1. CBS acknowledges support from the Philip Leverhulme Prize, the Royal Society Wolfson Fellowship, the EPSRC advanced career fellowship EP/V029428/1, EPSRC grants EP/S026045/1 and EP/T003553/1, EP/N014588/1, EP/T017961/1, the Wellcome Innovator Awards 215733/Z/19/Z and 221633/Z/20/Z, the European Union Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 777826 NoMADS, the Cantab Capital Institute for the Mathematics of Information and the Alan Turing Institute. The support from the personnel of the Institute of Diagnostic and Interventional Radiology and Nuclear Medicine, Wiener Neustadt, Austria, for the performance of measurements for the Alderson head phantom is gratefully appreciated.

Table 1: List of iterative methods detailed in the paper. Here 'Objective' describes the optimization problem that is solved: LS referring to the least-squares problem (3), Tikh. referring to the Tikhonov-regularized least-squares problem (7), hybrid referring to adding Tikhonov regularization to the projected problem at each iteration (9), TV referring to the least-squares problems with added total variation regularization (14). Note that Tikhonov regularization for a known regularization parameter λ can also be applied considering the augmented system (8) and using any solver for LS. Finally, note that algorithms cited with an asterisk required a significant adaptation from the original papers.

Method	**Description**	**Objective**	**Ref.**
CGLS	Conjugate gradient method applied to the normal equations. Minimizes the residual norm.	LS	[17]
LSQR	Mathematically equivalent to CGLS, using GK bidiagonalization, implemented with short recursions. Minimizes the residual norm.	LS	[36]
LSMR	Algortihm based on the GK bidiagonalization, minimizes the normal equations residual norm. The regularization parameter λ can be provided ahead of the iterations.	LS (λ=0) Tikh. ($\lambda\neq 0$)	[37]
AB-GMRES BA-GMRES	Adaptations of GMRES (minimal residual method using Arnoldi decomposition) using a given approximation of the backprojector as either left or right preconditioning. It is more robust for unmatched backprojectors.	LS	[27]
hybrid LSQR	Hybrid version of LSQR to solve Tikhonov regularized problems. The regularization parameter λ can be chosen ahead of the iterations or using a param. choice criteria (DP or GCV).	hybrid	[42]
TV-CGLS	Approximation of TV using a sequence of quadratic tangent majorants that are solved with CGLS.	TV	[44]*
TV-FLSQR	Approximation of TV using a sequence of quadratic tangent majorants that are partially solved throughout the iterations using FLSQR. It is faster than TV-CGLS but has a high storage cost.	TV	[48]*

Bibliography

[1] Miriam Leeser, Saoni Mukherjee, and James Brock. Fast reconstruction of 3D volumes from 2D CT projection data with GPUs. *BMC research notes*, 7(1):1–8, 2014.

[2] David C. Hansen and Thomas Sangild Sørensen. Fast 4d cone-beam ct from 60 s acquisitions. *Physics and Imaging in Radiation Oncology*, 5:69–75, 2018.

[3] Daniel Gulias-Soidan, Nilfa Milena Crus-Sanchez, Daniel Fraga-Manteiga, Juan Ignacio Cao-González, Vanesa Balboa-Barreiro, and Cristina González-Martín. Cone-beam ct-guided lung biopsies: Results in 94 patients. *Diagnostics*, 10(12), 2020.

[4] Ralph Kickuth, Claudia Reichling, Thorsten Bley, D. Hahn, and Carsten Ritter. C-arm cone-beam CT combined with a new electromagnetic navigation system for guidance of percutaneous needle biopsies: Initial clinical experience. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, volume 187, pages 569–576. © Georg Thieme Verlag KG, 2015.

[5] Ranald I. MacKay. Image guidance for proton therapy. *Clinical Oncology*, 30(5):293–298, 2018. Proton Beam and Particle Therapy.

[6] Sepideh Hatamikia, Ander Biguri, Gernot Kronreif, Michael Figl, Tom Russ, Joachim Kettenbach, Martin Buschmann, and Wolfgang Birkfellner. Toward on-the-fly trajectory optimization for c-arm cbct under strong kinematic constraints. *PLOS ONE*, 16(2):1–17, 02 2021.

[7] Mareike Thies, Jan-Nico Zäch, Cong Gao, Russel. H. Taylor, Navab Nassir, Andreas. K. Maier, and Mathias Unberath. A learning-based method for online adjustment of c-arm cone-beam ct source trajectories for artifact avoidance. *Int. J. Comput. Assist. Radiol. Surg*, 15:1787–1796, 2020.

[8] Avinash C. Kak and Malcolm Slaney. *Principles of computerized tomographic imaging*. SIAM, 2001.

[9] Per Christian Hansen. *Discrete Inverse Problems*. SIAM, 2010.

[10] Per Christian Hansen, Jakob Jørgensen, and William R. B. Lionheart. *Computed Tomography: Algorithms, Insight, and Just Enough Theory*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.

[11] Jennifer L. Mueller and Samuli Siltanen. *Linear and Nonlinear Inverse Problems with Practical Applications*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2012.

[12] Lee A. Feldkamp, Lloyd C. Davis, and James W. Kress. Practical cone-beam algorithm. *Josa a*, 1(6):612–619, 1984.

[13] Bharti Kataria, Jonas Nilsson Althén, Örjan Smedby, Anders Persson, Hannibal Sökjer, and Michael Sandborg. Assessment of image quality in abdominal CT: potential dose reduction with model-based iterative reconstruction. *European radiology*, 28(6):2464–2473, 2018.

[14] Weihua Mao, Chang Liu, Stephen J. Gardner, Farzan Siddiqui, Karen C. Snyder, Akila Kumarasiri, Bo Zhao, Joshua Kim, Ning Winston Wen, Benjamin Movsas, et al. Evaluation and clinical application of a commercially available iterative reconstruction algorithm for CBCT-based IGRT. *Technology in cancer research & treatment*, 18:1533033818823054, 2019.

[15] Gaurav S. Desai, Raul N. Uppot, Elaine W. Yu, Avinash R. Kambadakone, and Dushyant V. Sahani. Impact of iterative reconstruction on image quality and radiation dose in multidetector CT of large body size adults. *European radiology*, 22(8):1631–1640, 2012.

[16] Charalambos Rossides, Hossein Towsyfyan, Ander Biguri, Hans Deyhle, Reuben Lindroos, Mark Mavrogordato, Richard Boardman, Wenjuan Sun, and Thomas Blumensath. Effects of fast x-ray cone-beam tomographic measurement on dimensional metrology. *Metrologia*, 59(4):044003, 2022.

[17] Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards*, 49:409–435, 1952.

[18] Julianne Chung and Silvia Gazzola. Computational methods for large-scale inverse problems: a survey on hybrid projection methods, 2021.

[19] Silvia Gazzola and Malena Sabaté Landman. Krylov methods for inverse problems: Surveying classical, and introducing new, algorithmic approaches. *GAMM-Mitteilungen*, page e202000017, 2020.

[20] Andrei Dabravolski, Kees Joost Batenburg, and Jan Sijbers. Dynamic angle selection in x-ray computed tomography. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 324:17–24, 2014.

[21] Thanyawee Pengpen and Manuchehr Soleimani. Motion-compensated cone beam computed

tomography using a conjugate gradient least-squares algorithm and electrical impedance tomography imaging motion data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2043):20140390, 2015.

[22] Manasavee Lohvithee, Wenjuan Sun, Stephane Chretien, and Manuchehr Soleimani. Ant colony-based hyperparameter optimisation in total variation reconstruction in x-ray computed tomography. *MDPI Sensors*, 21(2):591, 2021.

[23] Daniil Kazantsev, Geert Van Eyndhoven, William R. B. Lionheart, Phillip J. Withers, Kate J. Dobson, Scott A. McDonald, Robert Atwood, and Peter D. Lee. Employing temporal self-similarity across the entire time domain in computed tomography reconstruction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2043):20140389, 2015.

[24] Monica Chillaron, Vicente Vidal, and Gumersindo Verdu. Evaluation of image filters for their integration with LSQR computerized tomography reconstruction method. *Plos one*, 15(3):e0229113, 2020.

[25] Liubov Flores, Vicente Vidal, Estíbaliz Parcero, and Gumersindo Verdú. Application of a modified LSQR method for CT imaging reconstruction with low doses to patient. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1969–1974. IEEE, 2016.

[26] Sophia B. Coban and William R. B. Lionheart. Regularised GMRES-type methods for x-ray computed tomography. *Technical report, University of Manchester*, 2014.

[27] Per Christian Hansen, Ken Hayami, and Keiichi Morikuni. GMRES methods for tomographic reconstruction with an unmatched back projector. *Journal of Computational and Applied Mathematics*, 413:114352, 2022.

[28] Emil Y. Sidky, Per Christian Hansen, Jakob S. Jørgensen, and Xiaochuan Pan. Iterative image reconstruction for CT with unmatched projection matrices using the generalized minimal residual algorithm. In Joseph Webster Stayman, editor, *7th International Conference on Image Formation in X-Ray Computed Tomography*, volume 12304, page 1230406. International Society for Optics and Photonics, SPIE, 2022.

[29] Silvia Gazzola, Per Christian Hansen, and James G Nagy. Ir tools: a matlab package of iterative regularization methods and large-scale test problems. *arXiv preprint arXiv:1712.05602*, 2017.

[30] Ander Biguri, Manjit Dosanjh, Steven Hancock, and Manuchehr Soleimani. TIGRE: a MATLAB-GPU toolbox for CBCT image reconstruction. *Biomedical Physics & Engineering Express*, 2(5):055010, 2016.

[31] Jakob S Jørgensen, Evelina Ametova, Genoveva Burca, Gemma Fardell, Evangelos Papoutsellis, Edoardo Pasca, Kris Thielemans, Martin Turner, Ryan Warr, William RB Lionheart, et al. Core imaging library-part i: a versatile python framework for tomographic imaging. *Philosophical Transactions of the Royal Society A*, 379(2204):20200192, 2021.

[32] Wim Van Aarle, Willem Jan Palenstijn, Jeroen Cant, Eline Janssens, Folkert Bleichrodt, Andrei Dabravolski, Jan De Beenhouwer, K Joost Batenburg, and Jan Sijbers. Fast and flexible x-ray tomography using the astra toolbox. *Optics express*, 24(22):25129–25147, 2016.

[33] Vojtěch Kulvait and Georg Rose. Software implementation of the krylov methods based reconstruction for the 3d cone beam ct operator. *arXiv preprint arXiv:2110.13526*, 2021.

[34] Curtis R. Vogel. *Computational Methods for Inverse Problems*. SIAM, Philadelphia, PA, USA, 2002.

[35] Daniela Calvetti. Preconditioned iterative methods for linear discrete ill-posed problems from a bayesian inversion perspective. *Journal of Computational and Applied Mathematics*, 198(2):378–395, 2007. Special Issue: Applied Computational Inverse Problems.

[36] Christopher C. Paige and Michael A. Saunders. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, 8:43–71, 1982.

[37] David Chin-Lung Fong and Michael Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM J. Sci. Comput.*, 33(5):2950–2971, 2011.

[38] Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[39] Christopher C. Paige and Michael A. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975.

[40] Ander Biguri, Reuben Lindroos, Robert Bryll, Hossein Towsyfyan, Hans Deyhle, Ibrahim El khalil Harrane, Richard Boardman, Mark Mavrogordato, Manjit Dosanjh, Steven Hancock, and Thomas Blumensath. Arbitrarily large tomography with iterative algorithms on multiple GPUs using the TIGRE toolbox. *Journal of Parallel and Distributed Computing*, 146:52–63, 2020.

[41] Julianne Chung and Katrina Palmer. A hybrid LSMR algorithm for large-scale Tikhonov regularization. *SIAM J. Sci. Comput.*, 37(5):S562–S580, 2015.

[42] Christopher C. Paige and Michael A. Saunders. Algorithm 583: LSQR: Sparse linear equations and least squares problems. *ACM Trans. Math. Softw.*, 8(2):195–209, jun 1982.

[43] Vladimir Alekseevich Morozov. On the solution of functional equations by the method of regularization. *Soviet Math. Dokl.*, 7:414–417, 1966.

[44] Brendt Wohlberg and Paul Rodriguez. An iteratively reweighted norm algorithm for minimization of total variation functionals. *Signal Processing Letters, IEEE*, 14:948–951, 01 2008.

[45] Simon R. Arridge, Marta M. Betcke, and Lauri Harhanen. Iterated preconditioned LSQR method for inverse problems on unstructured grids. *Inverse Problems*, 2014.

[46] Daniela Calvetti. Preconditioned iterative methods for linear discrete ill-posed problems from a Bayesian inversion perspective. *J. Comput. Appl. Math.*, 198(2):378–395, 2007. Special Issue: Applied Computational Inverse Problems.

[47] Silvia Gazzola and Malena Sabaté Landman. Flexible GMRES for total variation regularization. *Bit Numer. Math.*, 59:721–746, 2019.

[48] Silvia Gazzola, Sebastian J. Scott, and Alastair Spence. Flexible Krylov methods for edge enhancement in imaging. *Journal of Imaging.*, 7:43–71, 2021.

[49] Jingyan Xu and Benjamin M.W. Tsui. Electronic noise modeling in statistical iterative reconstruction. *IEEE Transactions on Image Processing*, 18(6):1228–1238, 2009.

[50] Yan Liu, Jianhua Ma, Yi Fan, and Zhengrong Liang. Adaptive-weighted total variation minimization for sparse data toward low-dose x-ray computed tomography image reconstruction. *Physics in Medicine & Biology*, 57(23):7923, 2012.

[51] Emil Y. Sidky and Xiaochuan Pan. Image reconstruction in circular cone-beam computed tomography by constrained, total-variation minimization. *Physics in Medicine & Biology*, 53(17):4777, 2008.

[52] Xun Jia, Yifei Lou, John Lewis, Ruijiang Li, Xuejun Gu, Chunhua Men, William Y. Song, and Steve B. Jiang. GPU-based fast low-dose cone beam CT reconstruction via total variation. *Journal of X-ray science and technology*, 19(2):139–154, 2011.

[53] Bruno De Man and Samit Basu. Distance-driven projection and backprojection in three dimensions. *Physics in Medicine & Biology*, 49(11):2463, 2004.

[54] Richard Martin Huber. *Pixel-Driven Projection Methods' Approximation Properties and Applications in Electron Tomography*. PhD thesis, University of Graz, 2022.

[55] Jean-François Pambrun and Rita Noumeir. Limitations of the SSIM quality metric in the context of diagnostic imaging. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 2960–2963, 2015.

[56] B. Girod. Psychovisual aspects of image processing: What's wrong with mean squared error? In *Proceedings of the Seventh Workshop on Multidimensional Signal Processing*, pages P.2–P.2, 1991.

[57] Per Christian Hansen, Jakob Sauer Jørgensen, and Peter Winkel Rasmussen. Stopping rules for algebraic iterative reconstruction methods in computed tomography. In *2021 21st International Conference on Computational Science and Its Applications (ICCSA)*, pages 60–70, 2021.