# Compressive Computed Tomography Reconstruction through Denoising Approximate Message Passing

**Citation for published version:**
Perelli, A, Lexa, M, Can, A & Davies, ME 2020, 'Compressive Computed Tomography Reconstruction through Denoising Approximate Message Passing', *Siam journal on imaging sciences*, vol. 13, no. 4, pp. 1860–1897. https://doi.org/10.1137/19M1310013

**Digital Object Identifier (DOI):**
10.1137/19M1310013



**Published In:**
Siam journal on imaging sciences

# Compressive Computed Tomography Reconstruction through Denoising Approximate Message Passing*

Alessandro Perelli[†], Michael Lexa[‡], Ali Can[‡], and Mike E. Davies[§]

**Abstract.** X-ray Computed Tomography (CT) reconstruction from a sparse number of views is a useful way to reduce either the radiation dose or the acquisition time, for example in fixed-gantry CT systems, however this results in an ill-posed inverse problem whose solution is typically computationally demanding. Approximate Message Passing (AMP) techniques represent the state of the art for solving undersampling Compressed Sensing problems with random linear measurements but there are still not clear solutions on how AMP should be modified and how it performs with real world problems. This paper investigates the question of whether we can employ an AMP framework for real sparse view CT imaging? The proposed algorithm for approximate inference in tomographic reconstruction incorporates a number of advances from within the AMP community, resulting in the Denoising Generalised Approximate Message Passing CT algorithm (D-GAMP-CT). Specifically, this exploits the use of sophisticated image denoisers to regularise the reconstruction. While in order to reduce the probability of divergence the (Radon) system and Poisson non-linear noise model are treated separately, exploiting the existence of efficient preconditioners for the former and the generalised noise modelling in GAMP for the latter. Experiments with simulated and real CT baggage scans confirm that the performance of the proposed algorithm outperforms statistical CT optimisation solvers.

**Key words.** X-ray Computed Tomography, Compressed Sensing, Approximate Message Passing, Image denoising, Preconditioning, Iterative algorithms

**AMS subject classifications.** 47A52, 49M30, 65J22, 94A08

## 1. Introduction.

X-ray Computed Tomography (CT) is one of the most widely used imaging techniques for medical diagnosis, image-guided radiotherapy, material characterization and security applications. Reducing X-ray radiation exposure is an important concern in particular for diagnostic CT where patients are subjected to repeated scans and for CT baggage scanners since the transmitted energy is related to the lifetime of the X-ray source. Furthermore, CT scanners employing Dual Energy (DE) systems tend to either reduce the acquisition data per energy or increase the dose and acquisition time. To lower the X-ray dose, two different strategies can be implemented: reducing the X-ray flux toward each detector element, i.e. the milliampere per seconds (low-mAs) per projection, or decrease the number of projections (sparse-views) per rotation. Similarly fixed gantry systems, e.g. [26], designed to accelerate scan time tend to further restrict the set of projections that can be acquired.

*Submitted to the editors of SIIMS.

**Funding:** This work was supported by U.S. Department of Homeland Security, Science and Technology Directorate, Explosives Division, BAA 13-05, Contract # HSHQDC-14-C-B0048, EPSRC platform grant EP/J015180/1 and the ERC Advanced grant, project C-SENSE, (ERC-ADG-2015-694888). A. Perelli has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 713683 (COFUNDfellowsDTU).

[†]Technical University of Denmark, DTU, Lyngby, DK (alper@dtu.dk).

[‡]GE Research Center, Niskayuna, US (lexa@ge.com, can@ge.com).

[§]Institute for Digital Communications (IDCOM), The University of Edinburgh, UK (m.davies@ed.ac.uk)

CT image reconstruction from sparse views and low dose, achieved by conventional filtered back projection (FBP) algorithms, is generally affected by noticeable streaking artifacts, due to insufficient sampling, and is not of acceptable quality for diagnostic purposes [8]. There is therefore a need in CT imaging applications for high quality image reconstruction algorithms that can accommodate sparse views and low dose. Many approaches have been proposed to solve this problem [62]. In particular, state of the art statistical image reconstruction typically aims to minimize a cost function defined as a sum of a data fidelity term that takes into account the measurement's statistical model and the geometry of the acquisition system, and a regularization term that imposes a prior model on the solution. Generally, the cost function for X-ray CT is either the negative log-likelihood function [18] or a penalized weighted least-squares (PWLS) cost function with a weighted quadratic approximation of the Poisson measurement noise model [28], [38]. Although several types of iterative algorithms have been designed to solve the statistical X-ray CT problem which can provide images with enhanced resolution and reduced artifacts compared to the FBP [57], in general current methods require many iterations to converge yielding a high computation time, and are often not suitable for clinical/industrial CT uses [6].

A large number of iterative algorithms have been utilized for statistical CT reconstruction, among these are coordinate descent [73], preconditioned conjugate gradient [22] and ordered subsets [19]. Recently researchers have developed new algorithms with faster convergence by using splitting techniques [47], alternating direction method of multipliers based algorithm [12] or combining Nesterov momentum techniques with ordered subsets to accelerate gradient descent methods [31]. In general, any first-order iterative method requires at each iteration the computation of at least one forward and back projection operator, together with a proximal mapping to account for the regularization term. These represent the main contributions to the overall computational time. In order to accelerate the reconstruction, it is therefore necessary to either design faster CT operators or develop iterative algorithms that can converge in fewer iterations.

In this work, we investigate the use of an emerging reconstruction method from Compressed Sensing (CS), called Approximate Message Passing (AMP) [17], for sparse view CT reconstruction. AMP based inference refers to a family of iterative algorithms first proposed in [17] for Compressed Sensing problems with an i.i.d. random Gaussian system matrix and a sparse signal model. AMP is a form of approximate Bayesian inference based on the notion of message passing or loopy belief propagation and is also strongly connected to the family of Expectation Propagation and Expectation Consistent approximation algorithms [42]. In essence, message passing algorithms work by iteratively updating marginal probabilities on the unknown variables until a locally consistent posterior probability model is obtained. The compelling aspect of the AMP family of algorithms is that they are designed to work in the large system limit (for random systems) which enables the central limit theorem to be invoked. This in turn simplifies the messages to be Gaussian distributions, requiring the algorithm to only pass means and variances. The result is a very efficient algorithm that is remarkably similar to the more traditional iterative shrinkage algorithm but with an additional "Onsager correction term" [17]. It also has many similarities to the Alternating Direction Method of Multipliers (ADMM) algorithm [51].

Today, AMP based algorithms provide the state-of-the-art performance in CS reconstruction both in terms of computation and reconstruction performance, e.g. [69, 27, 35]. For Gaussian measurements and random i.i.d. Gaussian sensing matrix, the convergence of the algorithm has been theoretically proven and, furthermore, it can be accurately quantified through its state evolution (SE) equations, exhibiting exponential convergence, in practice converging in very few iterations, for soft-thresholding function [5], linear denoisers [4] and a class of non-separable denoisers [7]. AMP can also incorporate non-Gaussian noise models through Rangan's Generalized Approximate Message Passing (GAMP) [50] and can approximate the Minimum Mean Square Error (MMSE) estimator by using a correctly matched prior or by exploiting learning structures such as Expectation Maximization [69] or SURE [27]. It has even been shown to be capable of incorporating sophisticated black box denoising algorithms in place of a signal prior model, resulting in the Denoising AMP (D-AMP) framework [35]. However, a key criticism directed at AMP, and its generalizations, is that they are specialist algorithms for i.i.d. and related measurement matrices and hence it is unclear to what extent they can be successfully applied to real word sensing problems.

There has been some work exploring the convergence properties of AMP and its generalizations to other matrix classes [53], S-AMP [10], and linking the algorithm with more classical optimization strategies such as ADMM [51]. A key problem of AMP is that when the measurement matrix is poorly conditioned and/or contains a significant mean offset the algorithm tends to diverge. One strategy for tackling this that is commonly used in loopy belief propagation is to incorporate damping to help stabilize the algorithm, [59, 68]. However, damping comes at the cost of significantly reducing the algorithm's convergence speed. It is also not clear what the value of the Onsager term is for general (deterministic) measurement matrices and whether the SE equations still provide a good prediction of the algorithm's performance.

Finally, Vector AMP [52] is a class of convergent algorithms with poorly conditioned matrices but the behavior with Fourier matrices is not rigorously understood. In summary, VAMP is a promising method for Fourier-based imaging while we do not know whether AMP based techniques can provide a competitive reconstruction framework to state-of-the-art methods for general real world imaging problems. The aim of this paper is to explore these issues for the specific case of sparse view CT imaging.

### 1.1. Main Contributions.

Our approach to develop an AMP based algorithm for CT reconstruction builds on a number of the recent developments in the field and, in particular, it makes use of the following key points: i) the design of a good preconditioner for the system based on the forward measurement model; ii) the inclusion of a non-linear Poisson noise model through the GAMP formulation; and iii) the incorporation of a broader class of signal prior than sparsity based models, through the D-AMP framework to enable the exploitation of state-of-the-art image denoising functions. We also demonstrate empirically the value of the Onsager term in the resulting algorithm and the accuracy of the generalized state evolution equations [50] even in this non-random setting.

As far as the authors are aware, this is the first work aimed at designing a denoising message passing based algorithm for CT reconstruction. A key challenge in applying GAMP to CT is the fact that the CT measurement operator for parallel or fan beam geometry has the form of a Radon type transform and is very ill-conditioned. This would require a significant level of damping to stabilize it and would be extremely slow [53]. The solution that we follow here is to replace the ill-conditioned operator with a much better conditioned one through preconditioning, exploiting the filtered back projection property of the system model [37]. The same procedure can be applied for different CT geometries like 2D fan-beam and 3D helical.

Another key challenge for CT reconstruction is how to accurately represent the Poisson noise model in the system. This can be approximated as a weighted $L_2$ error criterion [19], but then the preconditioner needs to account for both the system operator and the weighting matrix. While such preconditioners have been proposed, e.g. [31], they do not exploit the geometry of the measurement system and the subsequent system remains poorly conditioned, resulting in only modest improvements in convergence. In contrast, we will see that in the GAMP framework [50] the system operator and the noise process are naturally decoupled. This allows us a fully exploit a geometric preconditioner [37].

The final ingredient of our algorithm, which we call D-GAMP-CT, is the incorporation of the Block-matching and 3D filtering (BM3D) denoiser [14] to implicitly define a signal model through a sophisticated denoiser, rather than simply a sparse factorizable prior distribution [35]. While the proposed approach can leverage generic denoisers, we have utilized the BM3D denoiser since it provides state-of-the-art accuracy performance among deterministic denoisers and it also exploits a new implementation with reduced computational complexity [44]. We note that this framework is not restricted to the use of the BM3D denoiser but can be further extended to deep learning-based denoisers [74]. However for deep learning-based denoisers, it is crucial for achieving high denoising performance to have a high quality noiseless training database and it is often challenging or infeasible to obtain noiseless images in medical imaging.

We will see that the flexibility of using such a denoiser within GAMP yields to a better reconstruction of the image structure compared to more popular regularization, such as Total Variation (TV) minimization.

### 1.2. Relation to Existing Work.

The main issue of stabilizing AMP algorithms for non i.i.d. measurement matrices has already received attention in the literature. As previously discussed, damping is a popular solution [53, 59, 68] and, for example, has been applied successfully to hyperspectral imaging reconstruction [70, 64]. Schemes have also been proposed for modifying the algorithm when the matrix contains a significant non-zero offset [68, 33]. These approaches are fundamentally different from the one we present here where both issues are solved through our choice of a geometric preconditioner. Other aspects of our algorithm, such as the exploitation of general denoisers [35, 7], and the use of generalized noise models [50] have already appeared in the literature. Here we combine these to define a state of the art algorithm for sparse view CT reconstruction.

A new class of AMP algorithms called Vector AMP (VAMP) [52] (and the similar orthogonal AMP in [32]) that directly tackle the ill-conditioning problem in AMP by exploiting the singular value decomposition (SVD) of the measurement matrix. Such algorithms exhibit impressive performance and have provable reconstruction guarantees for the class of right-orthogonally invariant random matrices characterized by a scalar SE equation. The main intuition for such algorithms is that using the SVD of the measurement matrix , $\mathbf{\Phi} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, the right-orthogonal random component, $\mathbf{V}^T$ can be decoupled from the poorly conditioned component, $\mathbf{US}$ which is dealt with via a linear MMSE estimator component within the VAMP iteration [52]. While this significantly increases the class of matrices for which AMP techniques can be applied it still requires the calculation of the SVD. For large imaging problems, such as 2D or 3D CT imaging such a calculation is not practical as the operators themselves are computed on the fly and not stored in matrix form. In contrast, the approach we propose here similarly removes the ill-conditioning, but by right-multiplying by an easy to compute preconditioner, thus making it more attractive to large scale CT imaging applications. Another difference from VAMP is therefore that our preconditioner modifies the signal space and thus the signal model is defined in the preconditioned space rather than the original image space.

Finally, it is useful to draw a link with the existing literature on model based iterative reconstruction (MBIR) for CT imaging. Current state-of-the-art MBIR solutions for CT are based on minimizing a regularized negative log-likelihood (NLL) cost function or its approximation using penalized weighted least squares, [19, 18, 39], which can be interpreted as a Bayesian maximum a posteriori (MAP) estimator. This MAP framework can also be modified to incorporate denoising functions using the Plug-and-play (PnP) framework [48, 67]. In contrast, our proposed algorithm takes the MMSE estimator perspective on AMP and we analyse the equations of the SE prediction associate with the MMSE formulation of GAMP. Furthermore, as MAP estimation reduces to an optimization problem, the conditioning effects of the noise and system models are intertwined such that typical preconditioning has only a limited benefit. Using a preconditioned GAMP framework allows us to decouple these two effects.

### 1.3. Notation.

Matrices or discrete operators and column vectors are written respectively in capital and normal boldface type, i.e. $\mathbf{A}$ and $\mathbf{a}$ to distinguish from scalars and continuous variables written in normal weight. $(\cdot^T)$ and $(\cdot^H)$ refer respectively to the transpose and the conjugate transpose of a matrix and $\mathbf{1}$ refers to a vector of ones. Non-random quantities and random realizations are not distinguished typographically while random variable are written with capital letters. The conditional probability density function of $\mathbf{y}$ given $\mathbf{x}$ is denoted alternatively by $p_{Y|X}(\mathbf{y}|\mathbf{x})$ or $p(\mathbf{y}|\mathbf{x})$. A Gaussian random variable $\mathbf{x}$ with vector mean $\mathbf{a}$ and isotropic variance $b$ is denoted by $\mathbf{x} \sim \mathcal{N}(\mathbf{a}, b\mathbf{I})$. $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{b}^T\mathbf{a}$ refers to the vectors inner product.

### 1.4. Structure of the Paper.

The remainder of this paper is structured as follows: Section 2 briefly describes the physical model of transmission X-ray CT from the continuous to discrete domain, introduces the Poisson non linear noise model and the approximations that lead to the Plug-and-play statistical CT reconstruction problem. The section concludes with a discussion on the effects of the system and noise models on the conditioning of the problem. Section 3 reviews the original AMP algorithm for CS reconstruction, while Section 4 presents the proposed D-GAMP-CT algorithm highlighting the innovations which consist in utilizing the preconditioning for the Radon operator and incorporating the non linear CT Poisson noise model. Furthermore, we show empirical results for the SE of D-GAMP-CT. Finally, in Sections 6 and 7 comprehensive results of D-GAMP-CT on a numerical phantom and experimental acquisitions of cargo luggage are shown together with a comparison of its performance with state-of-the-art algorithm for model-based CT reconstruction.

## 2. X-ray Computed Tomography Model.

**2.1. Continuous-to-discrete model.** X-ray CT produces images of attenuation coefficients of the object or patient being scanned. A typical geometry of a CT scanner involves an incoherent source of X-ray radiation and a detector array recording the intensity of the radiation exiting the object along a number of paths. If the intensity of the source of radiation, $I_0$, passing through the object is known, then Beer's law provides the expected intensity after transmission, $I_i$ of the $i$-th ray as:

$$I_i = I_0 e^{-\int_{L_i} \mu(\vec{\nu})dl} + \epsilon_i \tag{2.1}$$

where $\int_{L_i} \cdot dl$ is the line integral along $L_i$ which is the path of the $i$th ray through the object from the source to the detector, $\mu(\vec{\nu})$ is the spatial distribution of attenuation and $\epsilon_i$ models the scatter and other background noise in the $i$th measurement. Equation (2.1) assumes a monoenergetic X-ray source which does not usually hold in practice. However, a common effective strategy for dealing with this consists of applying a polychromatic-to-monochromatic source correction pre-processing step [72], and in the rest of the paper we will therefore assume that we have a monoenergetic source or that it has already been appropriately corrected.

To obtain a discrete model, we should approximate the continuous attenuation function, $\mu(\vec{\nu}) \in L_2(\mathbb{R}^2)$, here defined over the 2D domain, using a finite basis expansion:

$$\mu(\vec{\nu}) \approx \sum_{j=1}^{N} \mu_j b_j(\vec{\nu}) \tag{2.2}$$

where $\boldsymbol{\mu} = [\mu_1, \dots, \mu_N]^T$ is the vector of attenuation coefficients and $b_j(\vec{\nu})$ define the $N$ basis functions associated with a discrete sampling on a $\sqrt{N} \times \sqrt{N}$ Cartesian grid. According to the parameterization in Eq. (2.2), the line integral becomes a summation:

$$\int_{L_i} \mu(\vec{\nu})dl \approx \sum_{j=1}^{N} \mu_j \int_{L_i} b_j(\vec{\nu})dl = \sum_{j=1}^{N} a_{ij}\mu_j. \tag{2.3}$$

where $a_{ij}$ represents the $i,j$ element of the system matrix describing the line integral along the $i$-path from source through object at pixel position $j$ onto each detector. Repeating this over all lines defines the full view linear tomographic system matrix $\mathbf{A} = [a_{ij}]$, where we assume that a sufficient density of lines has been taken such that the operator, $\mathbf{A}$, is one-to-one and hence invertible on its range, e.g. [2]. The matrix $\mathbf{A}$ is constructed as an over-determined matrix of dimensions $J \times N$ where $J$ is the product between the number of detectors $N_{dec}$ and the number of projections $N_\theta$.

Considering the sparse view scenario, the sub-sampled CT operator can now be represented as the application of a row sub-selection operator, $\mathbf{S}$ of dimensions $M \times J$, to $\mathbf{A}$, such that the linear part of the measurement system can be described in matrix form by

$$\mathbf{\Phi} = \mathbf{SA} \in \mathbb{R}^{M \times N} \tag{2.4}$$

with an effective undersampling ratio given by $M/N$.

In the case of normal exposure, the transmitted photon flux, $I_i$, follows a Poisson distribution. Using the discrete parameterization, Eqs. (2.2) and (2.3), we obtain the following discrete generalized linear model:

$$Y_i \sim \text{Poisson}\left\{I_0 e^{-z_i} + \epsilon_i\right\},\ i = 1, \ldots, M \tag{2.5}$$

where $z_i$ represents the discrete (linear) projection of the $i$th ray such that, $\mathbf{z} = \mathbf{\Phi}\boldsymbol{\mu}$.

**2.2. Sparse view CT reconstruction.** The sparse view CT reconstruction problem aims to estimate the attenuation coefficients, $\boldsymbol{\mu}$, from the measurements $\mathbf{y} = [y_1, \ldots, y_M]^T$ subject to Eq. (2.5) and any additional regularization. The negative log-likelihood (NLL) function for (2.5) given $\mathbf{y}$ has the form [18]:

$$-L(\boldsymbol{\mu}) = \sum_{i=1}^{M} \left\{ y_i \log\left[I_0 e^{-[\mathbf{\Phi}\boldsymbol{\mu}]_i} + \epsilon_i\right] - \left[I_0 e^{-[\mathbf{\Phi}\boldsymbol{\mu}]_i} + \epsilon_i\right]\right\}. \tag{2.6}$$

In the case of high/normal exposure a common practice is to use a quadratic approximation of Eq. (2.6) which leads to a Weighted Least Squares (WLS) approximation [18] based on taking the logarithm of the data, $l_i = \log\left(\frac{I_0}{y_i - \epsilon_i}\right)$. This is equivalent to observing $\mathbf{z}$ corrupted with a data-dependent Gaussian noise, $\mathbf{e}$, with inverse covariance $\mathbf{W} = \text{diag}\left[\frac{(y_i-\epsilon_i)^2}{y_i}\right]$:

$$\mathbf{l} = \mathbf{z} + \mathbf{e} = \mathbf{\Phi}\boldsymbol{\mu} + \mathbf{e} \tag{2.7}$$

The NLL can then be approximated as:

$$-L(\boldsymbol{\mu}) \approx \text{const.} + \left(\mathbf{\Phi}\boldsymbol{\mu} - \mathbf{l}\right)^T \mathbf{W}\left(\mathbf{\Phi}\boldsymbol{\mu} - \mathbf{l}\right). \tag{2.8}$$

For low dosage the logarithm cannot be utilized since the argument may not be non-negative, therefore Eq. (2.6) has to be used.

**2.3. Conditioning in sparse view CT.** It is instructive to consider the issues in minimizing (2.8). Most popular reconstruction algorithms solve a regularized form of (2.8) to further incorporate prior information of the image to be reconstructed:

$$\min_{\boldsymbol{\mu} \in \mathbb{R}_+^N} \frac{1}{2}||\mathbf{y} - \mathbf{\Phi}\boldsymbol{\mu}||_{\mathbf{W}}^2 + \lambda P(\boldsymbol{\mu}) \tag{2.9}$$

with $P$ usually a convex and possibly non-smooth regularization function. Assuming (2.9) is convex, many first order methods, like FISTA, can be applied to solve the optimization problem. Furthermore, it is possible to integrate denoising priors, such as BM3D or deep learning-based denoisers into ADMM or other algorithms using the non-convex Plug-and-play PP-WLS framework. However the convergence rate of such methods is highly dependent on the conditioning of the problem which in turn is a function of the Lipschitz constant of the data fit term $L = \sigma_{max}(\mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi})$ where $\sigma_{max}$ is the maximum eigenvalue. A large value of $L$ requires the use of a small step-size to ensure stability and results in slow convergence.

If the weighting matrix $\mathbf{W} \propto \mathbf{I}$, we are faced with the challenge of finding a preconditioner for the system matrix $\mathbf{\Phi} = \mathbf{SA}$ and fortunately there exist good preconditioners for this scenario based on the geometry of the tomographic problem. For example, this has been used in [37] where solutions for the direct inversion of $\mathbf{A}$ through a filter back projection operator are exploited. Indeed, both $\mathbf{W}$ and $\mathbf{\Phi}^T\mathbf{\Phi}$ are separately easy to precondition.

However, together, as in the PP-WLS framework, it is much more challenging. One approach that has been proposed [31] is to construct a diagonal preconditioner, $\mathbf{D}$, that majorizes the matrix, $\mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi}$:

$$\mathbf{D} = \mathrm{diag}\left(\mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi}\mathbf{1}\right) > \mathbf{\Phi}^T\mathbf{W}\mathbf{\Phi} \tag{2.10}$$

This solution exploits the non-negativity property of the measurement matrix $\mathbf{\Phi}$. Unfortunately, this type of preconditioner does not take into account the geometric structure in the system and therefore typically only provides modest speed improvements. Moreover adaptive methods for estimating the Lipschitz constant of accelerated first order solvers for composite minimization [25], backtracking line search [11] or adaptive restart [43] have been proposed. For the problem of CT reconstruction, heuristic line search techniques have been used as in [29]. However, these methods do not fundamentally change the Lipschitz constant and so may still be limited by ill-conditioning.

We will see that the GAMP framework enables us to avoid such problems by decoupling the measurement and noise components of the system. We are therefore able to exploit a preconditioner designed specifically for $\mathbf{A}$ which we detail next.

**2.4. Preconditioning of the Radon operator.** The aim is to replace the poorly conditioned operator, $\mathbf{A}$ with a new operator, $\widetilde{\mathbf{A}}$, that has a small condition number, i.e. it is a nearly tight frame, by mapping to a preconditioned image space. For 2D CT with parallel projections or fan-beam with appropriate resampling, our proposed solution is to use a cone filter applied in the image domain that amplifies high spatial frequencies, as has previously been used to accelerate reconstruction convergence of Conjugate Gradient solver for PWLS [22], [47](Sec. III D). In order to construct a discrete preconditioner, while staying geometrically exact to the continuous setting we follow the work [13]; since the operator $\mathbf{A}^T\mathbf{A}$ is approximately block-Toeplitz for shift invariant imaging CT problem, circulant preconditioners also called "Fourier" diagonalizing preconditioners have been applied to both image restoration problems [36](Sec. III A), and also shift-variant CT problems [22].
The continuous 2D X-ray Transform is a linear operator $\mathcal{A} : L_2(\mathbb{R}^2) \to L_2([0,\pi) \times \mathbb{R})$ which computes the line integral of a function in the 2D input space. The Fourier central slice Theorem states that

$$\mathcal{A} = \mathcal{F}_\gamma^{-1}\Omega_{\omega^{-1}}\mathcal{F} \tag{2.11}$$

where $\mathcal{F}$ is the 2D Fourier transform (FT), $\Omega$ is the coordinate transform operator from Cartesian to polar coordinates, $\omega^{-1} = (\gamma\cos\delta, \gamma\sin\delta)$ and $\mathcal{F}_\gamma^{-1}$ is the inverse 1D FT with respect to $\gamma$. The output of the linear operator is the sinogram that is a function of $\delta$ and the polar space variable $\rho$. Both $\mathcal{A}$ and $\mathcal{A}^T\mathcal{A}$ are normal-convolutional operators since

$$\mathcal{A}^T\mathcal{A}\mu = \mu \circledast \frac{1}{|\gamma|} \tag{2.12}$$

and in the frequency domain

$$\mathcal{A}^T\mathcal{A} = \mathcal{F}^H(\Omega_{\omega^{-1}})^T(\mathcal{F}_\gamma^{-1})^H\mathcal{F}_\gamma^{-1}\Omega_{\omega^{-1}}\mathcal{F} \overset{(a)}{=} \mathcal{F}^H(\Omega_{\omega^{-1}})^T\Omega_{\omega^{-1}}\mathcal{F}$$
$$\overset{(b)}{=} \mathcal{F}^H|\mathrm{det}J_\omega|\mathcal{F} = \mathcal{F}^H\mathcal{D}\left(\frac{1}{|\rho|}\right)\mathcal{F} \tag{2.13}$$

where (a) follows from $\mathcal{F}_\gamma$ being an unitary operator and (b) derives from the back-projection CT filter formulation where $J_\omega$ defines the Jacobian of $\omega$ and $\mathcal{D}\left(\frac{1}{|\rho|}\right)$ is the diagonal polar Fourier space operator. From the continuous to the discrete domain, the Fourier-based Radon transform can be written as [34, 40]:

$$\mathbf{A} = \mathbf{F}_\gamma^{-1}\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F} \tag{2.14}$$

where $\mathbf{F}$ is the 2D unitary discrete Fourier transform operator which takes as input the image $\boldsymbol{\mu}$ of dimensions $I \times I$, with $I = \sqrt{N}$. The operator $\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}$ performs a discretized version of the continuous coordinate transform in Eq. (2.12) which outputs a matrix of polar coordinate samples that are equally-spaced along $\rho$ at the discrete locations $\{i\Delta_\rho\}$ for $i = -\frac{I}{2}, \ldots, \frac{I}{2} - 1$. The degree of approximation between continuous to discrete domain within the Fourier-based approaches is determined by the non-uniform interpolation in the frequency space [45]. In [23], a min-max analysis provides the interpolator that minimizes the worst case error. Whilst no analytical formula exists for specifying the optimal choice of the scaling function, the Kaiser-Bessel interpolation kernel can provide good compromise between accuracy and simplicity.

The non-uniform FFT operator $\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F}$ takes as input the $I \times I$ input image matrix and output a matrix of dimensions $N_{dec} \times N_\theta$; $\mathbf{F}_\gamma$ applies the 1D unitary discrete Fourier transform (DFT) matrix separately to each of the radial lines vectors of dimension $N_{dec}$ and it is defined as the Kronecker product between a 1D DFT matrix $\mathbf{F}_1$ and the identity matrix, i.e. $\mathbf{F}_\gamma = \mathbf{F}_1 \otimes \mathbf{I}_{N_\theta}$. Therefore, the final output is a vector of dimensions $J = N_{dec} \cdot N_\theta$ where $N_{dec}$ is the number of detectors and $N_\theta$ is the number of angles (or number of projections) in agreement with (2.4). This formulation has the advantage of being approximately, up to the gridding interpolation, one-to-one. Since $\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F}$ and hence $\mathbf{A}$ are poorly conditioned, one solution is to replace $\mathbf{A}$ with a better conditioned modified transform $\widetilde{\mathbf{A}}$. This is equivalent to working in a new preconditioned signal space, $\mathbf{x} = \mathbf{V}\boldsymbol{\mu}$ via real-valued the linear transform

$$\mathbf{V} = \mathbf{F}^{-1}\mathbf{C}^{\frac{1}{2}}\mathbf{F} \tag{2.15}$$

where $\mathbf{C} = \mathrm{diag}\left(\frac{1}{\sqrt{a^2+b^2}}\right)$ is a diagonal matrix, on the vector space, that normalizes the FT components by the sampling rate relative to the Cartesian samples $(a, b)$, with $|\rho| = \sqrt{a^2 + b^2}$, which corresponds to the point spread function of the $\mathbf{A}^T\mathbf{A}$ at a one-pixel point source located at the center of the field-of-view [13](V.A). By applying $\mathbf{V}$ as right preconditioner for $\mathbf{A}$, we obtain the following expression

$$\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{V}^{-1} = \mathbf{F}_\gamma^{-1}\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F}\mathbf{F}^{-1}\mathbf{C}^{-\frac{1}{2}}\mathbf{F} = \mathbf{F}_\gamma^{-1}\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{C}^{-\frac{1}{2}}\mathbf{F} \tag{2.16}$$

It is worth noting that the operators defined in both Eqs. (2.14) and (2.16) and the linear transform $\mathbf{V}$ can be seen to remain real valued through the usual conjugate symmetry arguments. Since the operator $\mathbf{C}$ is symmetric and $\mathbf{F}_1$ is an orthogonal operator, the mapping from image to image results

$$\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}} = (\mathbf{A}\mathbf{V}^{-1})^T(\mathbf{A}\mathbf{V}^{-1}) = \mathbf{V}^{-T}(\mathbf{A}^T\mathbf{A})\mathbf{V}^{-1}$$
$$= \mathbf{V}^{-T}[\mathbf{F}^H(\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}})^T(\mathbf{F}_\gamma^{-1})^H\mathbf{F}_\gamma^{-1}\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F}]\mathbf{V}^{-1} \stackrel{(a)}{=} \mathbf{V}^{-T}[\mathbf{F}^H(\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}})^T\boldsymbol{\Omega}_{\boldsymbol{\omega}^{-1}}\mathbf{F}]\mathbf{V}^{-1}$$
$$\stackrel{(b)}{=} \mathbf{F}^{-1}\mathbf{C}^{-\frac{1}{2}}\mathbf{F}\left[\mathbf{F}^H\mathbf{C}\mathbf{F}\right]\mathbf{F}^{-1}\mathbf{C}^{-\frac{1}{2}}\mathbf{F} = \mathbf{I} \tag{2.17}$$

where (a) comes from $\mathbf{F}_\gamma$ being 1D unitary discrete Fourier transform matrix and (b) follows from the linear transformation in Eq. (2.15). The operator $\widetilde{\mathbf{A}}^T\widetilde{\mathbf{A}}$ is real, symmetric and positive definite. Both the matrix preconditioner in Eq. (2.15) and its inverse have fast $\mathcal{O}(I \log I)$ implementations. Other Fourier based preconditioners could have been chosen like the Pseudo Polar FT based left preconditioner. While it has the advantage that the operator is assured to be one-to-one and empirically the singular value spread of $\widetilde{\mathbf{A}}$, the left preconditioning (in the measurement space) changes the statistical noise model.

For sparse view CT, the row sub-sampling operator $\mathbf{S} \in \mathbb{R}^{M\times J}$ is applied, such that the overall linear measurement system can be expressed by

$$\widetilde{\boldsymbol{\Phi}} = \mathbf{S}\widetilde{\mathbf{A}} \in \mathbb{R}^{M\times N}. \tag{2.18}$$

An important consequence of applying such preconditioning is that the image prior to be used in the GAMP reconstruction framework needs to be defined on $\mathbf{x}$ in the preconditioned space. It will also be necessary to apply a final post-processing step to map the estimated vector, $\mathbf{x}$, back into the image domain $\boldsymbol{\mu}$.

## 3. Review of the Generalized Approximate Message Passing algorithm.

In this Section, we review the formulation of the GAMP algorithm proposed in [50] which is a generalization of the original AMP algorithm [17]. AMP belongs to a families of iterative algorithms for solving linear systems of the type in Eq. (2.8) based on different Gaussian approximations of loopy Belief Propagation. In this respect AMP, S-AMP, VAMP represent alternative ways to perform variational inference but all of them enjoy rigorous state evolution behavior. All these algorithms share the same iterative structure of performing a MAP or MSE estimation of the vector mean and scalar variance in the image domain and in the measurement domain.

$$\begin{array}{cc} \text{MAP or MSE} - \text{input domain} & \text{MAP or MSE} - \text{measurement domain} \\ \mathbf{x} = g_{in}(\mathbf{r}, \tau_r) & \mathbf{s} = g_{out}(\mathbf{p}, \tau_p) \\ \tau_x = \tau_r g'_{in}(\mathbf{r}, \tau_r) & \tau_s = -g'_{out}(\mathbf{p}, \tau_p) \end{array}$$

The difference between the algorithms relies on how the mean and variances are computed, i.e. the functions $g_{in}$, $g_{out}$ together with the vectors $\mathbf{r}, \mathbf{p}$ and to which classes of random measurement matrices they can be applied. We develop our framework based on the GAMP formulation which is detailed in the following and we will describe how it differs from the VAMP algorithm. GAMP considers a class of generalized linear Bayesian inference problems, precisely estimating an unknown high dimensional input vector $\boldsymbol{\mu} \in \mathbb{R}^N$ observed by a mixing random linear operator $\boldsymbol{\Phi} \in \mathbb{R}^{M\times N}$ followed by a component-wise and nonlinear noise measurement model.

In detail, the Bayesian forward model consists of an unknown random vector $\boldsymbol{\mu}$ generated from a prior separable distribution $p(\boldsymbol{\mu}) = \prod_{i=1}^{N} p(\mu_i)$; the input vector is then multiplied by a measurement matrix $\boldsymbol{\Phi}$ whose elements are i.i.d. random Gaussian distributed $\mathcal{N}\left(0, \frac{1}{M}\right)$, i.e. $\mathbf{z} = \boldsymbol{\Phi}\boldsymbol{\mu}$. Finally each component of the vector $\mathbf{z}$ generates a nonlinear output $y_j$, $j = 1, \ldots, M$ described by a conditional probability distribution (or likelihood) $p_{y|z}(\mathbf{y}|\mathbf{z})$.

Given the fully connected graphical model with arbitrary separable prior and separable likelihood, GAMP is an efficient and tractable message passing method based on a Gaussian approximation of loopy belief propagation (BP) in the large system limit. GAMP is constructed as an iterative algorithm which sequentially estimates the vector mean associated with samples $\boldsymbol{\mu}$ and $\mathbf{z}$ and the scalar second order statistic (variances). By construction, GAMP can perform Max-Sum loopy BP for approximate MAP estimation, or Sum-Product loopy BP computing approximate MMSE estimates; we will focus on the latter estimation problem in this paper.

GAMP algorithm converts the vector MMSE estimation problem to a sequence, indexed by $t$, of scalar MMSE estimations in the input signal and measurement domain, based on the large system limit assumption. Algorithmically, given the linear estimate $\mathbf{z}^t = \boldsymbol{\Phi}\boldsymbol{\mu}^t$, GAMP employs a MMSE estimator of $\mathbf{z}^t$, which results from a Gaussian approximation of the sum-product loopy BP on the dense graph (induced by $\boldsymbol{\Phi}$), and it propagates these means and isotropic variances estimates backward through $\boldsymbol{\Phi}$ to give a noisy estimate for $\boldsymbol{\mu}$. Then, the algorithm performs a MMSE estimate of $\boldsymbol{\mu}$ and propagates it forwards onto the measurements again. In order to approximately perform sum-product loopy BP and to obtain the MMSE estimates, GAMP provides a framework to construct two scalar functions in the input and measurement domain, $g_{out}(\cdot)$ and $g_{in}(\cdot)$ respectively. We review how to construct the function $g_{out}(\cdot)$ in the measurement domain; we consider the conditional probability distribution

$$p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t) \propto e^{\log p_{Y|Z^t}(\mathbf{y}|\mathbf{z}^t)} e^{-\frac{1}{2\tau_p^t}(\mathbf{z}^t - \mathbf{p}^t)^T(\mathbf{z}^t - \mathbf{p}^t)} \tag{3.1}$$

which can be interpreted as the posterior density function of the random variable $\Xi^t \sim \mathcal{N}(\mathbf{p}^t, \tau_p^t \mathbf{I})$ with observation $Y \sim p_{Y|Z^t}(\mathbf{y}|\mathbf{z}^t)$ where $Z^t$ is a random variable associated with the linear estimate whose instance is $z^t$. By construction of the approximate sum-product loopy BP, the messages $\Xi^t \sim \mathcal{N}(\mathbf{p}^t, \tau_p^t \mathbf{I})$ are Gaussian with scalar variance and the mean is defined as a perturbed version of the linear estimate $\mathbf{z}^t$, i.e.

$$\mathbf{p}^t = \mathbf{z}^t - \tau_p^t \mathbf{s}^{t-1} \tag{3.2}$$

where the term (perturbation) $\tau_p^t \mathbf{s}^{t-1}$ represents the Onsager term. Given $p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)$, the approximate iterative BP for the MMSE problem is achieved by computing

$$\mathbf{z}_0^t := \mathbb{E}_{p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)}[\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t] \tag{3.3}$$

$$g_{out}(\mathbf{p}^t, \mathbf{y}, \tau_p^t) = \mathbf{s}^t := \frac{1}{\tau_p^t}(\mathbf{z}_0^t - \mathbf{p}^t) \tag{3.4}$$

where $\mathbf{z}_0^t$ is the MMSE estimate of $Z^t$ given $\Xi^t$. The variance $\tau_s^t$ is calculated as the average of the negative derivative of $g_{out}(p_i^t, y_i, \tau_p^t)$ respect to $p_i \, \forall i = 1, \dots, M$ as follows

$$\tau_{s_i}^t = -\frac{\partial}{\partial p_i} g_{out}(p_i^t, y_i, \tau_p^t) \overset{(a)}{=} \frac{1}{\tau_p^t}\left[1 - \frac{\mathrm{Var}(z_i^t | p_i^t, y_i, \tau_p^t)}{\tau_p^t}\right] \tag{3.5}$$

$$\tau_s^t = \frac{1}{M}\sum_{i=1}^{M} \tau_{s_i}^t$$

where the equality (a) follows from the derivation in [50, Appendix D]. The vector mean of the linear estimate $R \sim \mathcal{N}(\mathbf{r}^t, \tau_r^t \mathbf{I})$ in the input domain is

$$\mathbf{r}^t = \mathbf{x}^t + \tau_r^t \mathbf{\Phi}^T \mathbf{s}^t \tag{3.6}$$

Finally, to obtain an approximate MMSE vector mean and scalar variance estimates in input signal domain given $\mathbf{r}^t$, the function $g_{in}(\cdot)$ has to be constructed as follows

$$g_{in}(\mathbf{r}^t) = \boldsymbol{\mu}^{t+1} = \mathbb{E}[\boldsymbol{\mu} | \mathbf{r}^t, \tau_r^t] \tag{3.7}$$

$$\tau_\mu^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Var}(\mu_i | r_i^t, \tau_r^t)$$

In the next Section we focus on the main modifications we have introduced the the GAMP algorithm which concern how to include the preconditioning in the linear operator, how to calculate Eq. (3.1) for the case of non linear Poisson noise model and extend to non separable input signal models, i.e. how to calculate (3.7) without the explicit knowledge of the prior distribution of the unknown input signal.

## 4. D-GAMP-CT: Denoising CT with Poisson noise based AMP.

The proposed algorithm for CT reconstruction is built upon the GAMP framework with the following innovation: i) incorporate the preconditioner for the Radon operator, introduced in Section 2.4, such that the iterative algorithm is performed in the preconditioned space together with a new operator with a smaller condition number. Furthermore, the algorithm utilises the following properties: ii) exploit the GAMP formulation (3) for the non linear Poisson noise model in Eq. (2.5); iii) use a generic denoiser in the non linear step to capture the data-dependent structure of complex images [35]. The benefit of employing i) and ii) relies on the property of decoupling the measurements and noise components unlike the solution in (2.10).

### 4.1. Preconditioning of the measurement operator.

As described in Section 2.4, the Radon operator (2.14) can be preconditioned by using (2.15) such that the combined operator $\widetilde{\mathbf{A}}$ has a condition number considerably lower than $\mathbf{A}$ [1, 3]. Combining (2.15) and (2.4), we define the modified system matrix, $\widetilde{\mathbf{\Phi}}$, as

$$\widetilde{\mathbf{\Phi}} = \mathbf{S}\mathbf{A}\mathbf{V}^{-1} = \mathbf{\Phi}\mathbf{V}^{-1} \tag{4.1}$$

$$\widetilde{\mathbf{\Phi}}^T = \mathbf{V}^{-1}\mathbf{A}^T\mathbf{S}^T = \mathbf{V}^{-1}\mathbf{\Phi}^T$$

The computational complexity of both operators, $\widetilde{\mathbf{\Phi}}$ and $\widetilde{\mathbf{\Phi}}^T$ is of order $\mathcal{O}(N \log N)$, since they are defined as a composition of element-wise operators with complexity $\mathcal{O}(N)$ and the

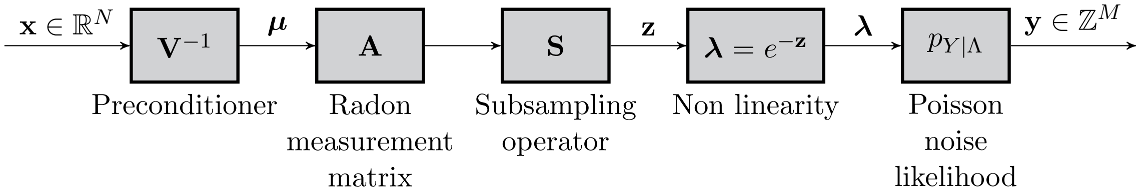



Figure 1: Computed Tomography estimation model with Poisson noise model and matrix preconditioner $\mathbf{V}$ in the image domain.

FFT, with complexity $\mathcal{O}(N \log N)$. In an equivalent way, the preconditioning leads to the following change of coordinates in the signal domain within each iteration $t$:

$$\boldsymbol{\mu}^t = \mathbf{V}^{-1}\mathbf{x}^t \quad \rightarrow \quad \mathbf{x}^t = \mathbf{V}\boldsymbol{\mu}^t$$

**4.2. Incorporation of the Poisson Noise Model in GAMP.** We consider the sparse views X-ray CT transmission model where the input vector $\boldsymbol{\mu} \in \mathbb{R}^N$ is passed through the linear Radon CT operator together with the angular subsampling operator, that is modelled as

$$\lambda_a = e^{-z_a} = e^{-[\widetilde{\mathbf{\Phi}}\mathbf{x}]_a}, \quad a = 1, \dots, M \tag{4.2}$$

where the linear term is $\mathbf{z} = \mathbf{S}\mathbf{A}\boldsymbol{\mu} = \mathbf{\Phi}\boldsymbol{\mu} = \widetilde{\mathbf{\Phi}}\mathbf{x}$ from Eq. (4.1) and (4.2). Finally, each component $\lambda_a$ randomly generates an output component $y_a$ of the vector $\mathbf{y} \in \mathbb{Z}^M$. The conditional probability distribution of the i.i.d. random variable $Y$ given the linear measurement $Z$ is an exponential-Poisson distribution [39]

$$p_{Y|Z}(\mathbf{y}|\mathbf{z}) = \prod_{a=1}^{M} \frac{1}{y_a!} e^{-(e^{-z_a})} e^{-y_a z_a} \tag{4.3}$$

Fig. 1 shows the block diagram of the generative measurement model; $\mathbf{x}$ is the preconditioned vector which is mapped through $\mathbf{V}^{-1}$ to the vector representing the CT attenuation coefficients $\boldsymbol{\mu}$. It is the input of the Radon system model $\mathbf{A}$ and subsequently to the sparse view operator $\mathbf{S}$ as described in Eq. (2.4). According to the transmission CT model an exponential non-linearity is applied to the corresponding linear measurement vector $\mathbf{z}$. Finally, the Poisson likelihood $p_{Y|Z}$ models the CT noise as described in Eq. (2.5), given the linear system $\mathbf{z} = \widetilde{\mathbf{\Phi}}\mathbf{x}$. While the expression of the likelihood $p_{Y|Z}$ in Eq. (4.3) relates the random variables $Y$ and $Z$, and therefore it already includes the non-linearity (4.2), in Fig. (1) we have highlighted the non-linear block followed by an auxiliary likelihood $p_{Y|\Lambda}$.

In this section, we describe how to perform the MMSE estimation in the measurement domain for the nonlinear CT Poisson noise model. Given $p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)$ as defined in Eq. (3.1) with $p_{Y|Z}(\mathbf{y}|\mathbf{z})$ in (4.3) and the vector $\mathbf{p}^t$ detailed in line (6) of the Algorithm 4.1, the approximate iterative BP for the MMSE problem is achieved by computing

$$\mathbf{s}^t(\mathbf{p}^t, \mathbf{y}, \tau_p^t) = \frac{1}{\tau_p^t}(\mathbf{z}_0^t - \mathbf{p}^t), \quad \mathbf{z}_0^t := \mathbb{E}_{p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)}[\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t] \tag{4.4}$$

To obtain $\mathbf{s}^t(\mathbf{p}^t, \mathbf{y}, \tau_p^t)$, we need to evaluate the expectation $\mathbb{E}(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)$ respect to $p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t)$, where

$$\frac{1}{M}\log p(\mathbf{y}|\mathbf{z}^t) = -\langle \mathbf{z}^t, \mathbf{y}\rangle - \langle e^{-\mathbf{z}^t}, \mathbf{1}\rangle - \langle \log(\mathbf{y}!), \mathbf{1}\rangle \tag{4.5}$$

$$p(\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}) \propto e^{-\langle \mathbf{z}^t,\mathbf{y}\rangle - \langle e^{-\mathbf{z}^t},\mathbf{1}\rangle - \langle \log(\mathbf{y}!),\mathbf{1}\rangle - \frac{1}{2\tau_p^t}||\mathbf{z}^t-\mathbf{p}^t||_2^2}, \ \mathbf{z}^t \in \mathbb{R}^M_{\geq 0}$$

The expectation requires solving the following ratio of integrals for each element indexed with $a = 1, \dots, M$:

$$\mathbb{E}[z_a^t|p_a^t, y_a, \tau_p^t] = \frac{\int_{\mathbb{R}_{\geq 0}} z_a^t e^{\log p_{Y|Z^t}(y_a|z_a^t)} e^{-\frac{1}{2\tau_p^t}(z_a^t - p_a^t)^2} dz_a^t}{\int_{\mathbb{R}_{\geq 0}} e^{\log p_{Y|Z^t}(y_a|z_a^t)} e^{-\frac{1}{2\tau_p^t}(z_a^t - p_a^t)^2} dz_a^t} \tag{4.6}$$

Unfortunately no closed form solution appears to exist and therefore Laplace's method [65] is used to approximate the posterior mean $\mathbf{z}_0^t$ and $\tau_s^t$. In Appendix A, the calculation for $\mathbf{z}_0^t$ and $\mathrm{Var}[\mathbf{z}^t|\mathbf{p}^t]$ is detailed.

It is worth noting that the solution obtained by BM3D-CT-GAMP, using the Poisson noise model, is different from the solution of the regularized NLL minimization problem stated in Eq. (2.6).

### 4.3. Denoising: Non-Linear Input Module.

Whilst the original GAMP algorithm was developed on a factorial (sparse) signal model, the framework has been shown to be amenable to much broader classes of estimators [35, 7]. Since the GAMP algorithm approximates the estimate for $\mathbf{x}$ as a Gaussian noise corrupted version of the true signal with variance $\tau_r^t$ as in Eq. (4.8), it is meaningful to employ, instead of a prior-based non linear scalar function, a denoiser $D_{\tau_r^t}$ which acts as a standard non-linear mapping

$$D_{\tau_r^t}(\cdot) : \mathbb{R}^N \to \mathbb{R}^N, \quad \mathbf{r} \longmapsto D_{\tau_r^t}(\mathbf{r}) \tag{4.7}$$

that, given a noisy signal estimate

$$\mathbf{r} = \mathbf{x} + \sqrt{\tau_r^t}\boldsymbol{\psi} \tag{4.8}$$

with $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, outputs an estimate of $\mathbf{x}$. We treat $D_{\tau_r^t}(\cdot)$ as a black box estimator, i.e., we do not require knowledge of its functional form [35].

The main reason for using a generic denoiser in the non linear step is to capture the data-dependent structure of complex images, rather than a simple factorial model, obtaining a sequence of estimates eventually converging faster to the true preconditioned signal $\mathbf{x}$; this provides the flexibility in using a variety of denoisers. Given the estimated signal

$$\mathbf{r}^t = \mathbf{x}^t + \tau_r^t \mathbf{V}^{-1}\boldsymbol{\Phi}^T\mathbf{s}^t \tag{4.9}$$

which is the input of the denoiser, the output vector estimate and the scalar variance are given by

$$\begin{aligned} \mathbf{x}^{t+1} &= D_{\tau_r^t}(\mathbf{r}^t) \\ \tau_x^{t+1} &= \tau_r^t D'_{\tau_r^t}(\mathbf{r}^t) \end{aligned} \tag{4.10}$$

where $D'_{\tau_r^t}(\cdot)$ denotes the divergence of the denoiser, which is by definition the sum of the partial derivatives with respect to each element $x_i, i = 1, \ldots, N$ of $\mathbf{x}$ and it is a scalar, i.e.

$$D'_{\tau_r^t}(\mathbf{r}) = \mathrm{div}_{\mathbf{r}}(D_{\tau_r^t}(\mathbf{r})) = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial D_{\tau_r^t}(\mathbf{r})}{\partial r_i} \tag{4.11}$$

For input vector $\mathbf{r}$ belonging to a simple class of signals $\mathcal{C}$, it is possible to construct a denoiser which compute the conditional mean as in the MMSE formulation of GAMP in Eq. (3.7), but for a general class of signals, the denoiser $D_{\tau_r^t}(\mathbf{r})$ does not necessarily correspond to a mean estimator. By design, the denoiser $D_{\tau_r^t}(\cdot)$ is acting in the preconditioned image space which consists of a high pass filtering of the image in the original spatial domain. The main properties of the denoiser are to be monotone, which means that the risk

$$R(\mathbf{x}, \tau_r^t) = \frac{1}{N}\mathbb{E}\|D_{\tau_r^t}(\mathbf{x} + \sqrt{\tau_r^t}\boldsymbol{\psi}) - \mathbf{x}\|_2^2 \tag{4.12}$$

is a non-decreasing function of $\tau_r^t$, and proper, i.e. $\sup_{\mathbf{x}\in\mathcal{C}} R(\mathbf{x}, \tau_r^t) \leq \nu\tau_r^t$, for $\nu \in (0, 1)$; this implies that given an estimate of $\tau_r^t$, it results $\tau_r^{t+1} \leq \tau_r^t$. Therefore, even when the input of the denoiser belongs to the preconditioned space, as in Eqs. (4.2) - (4.9) which highlight the fact that the noise is no more uncorrelated, traditional denoisers can still be used at the cost of a decrease in the rate or reduction of the risk. Similar arguments are used in plug-and-play framework [9] where at each iteration the noise term is in general correlated to the signal.

In our framework, because of the design of the preconditioner as a high-pass filter, it is possible to utilize better denoisers which can handle this signal mapping; one choice that we have compared in the result is by using a modification of the proximal-based TV denoiser where the $\|\cdot\|_{(\mathbf{V}^T\mathbf{V})^{-1}}$ norm is used instead of the $l_2$ norm.

In Section 6.3, we show that using this tailored denoiser leads to an improvement in the accuracy error only at earlier iterations, before convergence, compared to $l_2$ proximal TV map. The analytic calculation of $D'_{\tau_r^t}(\cdot)$ is often not available and it is in general data-dependent, but a good approximation can be obtained through the Monte Carlo technique. In [46] the authors showed that given a denoiser $D_{\tau_r^t}(\cdot)$ and an i.i.d. random vector $\mathbf{b} \sim \mathcal{N}(0, \mathbf{I})$, the divergence can be estimated as

$$D'_{\tau_r^t}(\mathbf{r}) \approx \frac{1}{N}\mathbb{E}_{\mathbf{b}}\left[\frac{1}{\epsilon}\mathbf{b}^T\Big(D_{\tau_r^t}(\mathbf{r} + \epsilon\mathbf{b}) - D_{\tau_r^t}(\mathbf{r})\Big)\right],\ \epsilon \to 0 \tag{4.13}$$

where the expectation over the random variable $\mathbf{b}$ is calculated using a Monte-Carlo method, i.e. generate $K$ i.i.d. $\mathcal{N}(0, \mathbf{I})$ samples vectors, estimate the divergence for each vector and then obtain the global divergence by averaging:

$$D'_{\tau_r^t}(\mathbf{r}) = \frac{1}{NK}\sum_{j=1}^{K}\mathbf{b}_j^T\left(\frac{D_{\tau_r^t}(\mathbf{r} + \epsilon\mathbf{b}_j) - D_{\tau_r^t}(\mathbf{r})}{\epsilon}\right) \tag{4.14}$$

Given the vectorized image lying in a high dimensional space, it has been empirically observed [46] that we can accurately approximate the expected value using only a single random sample, i.e. $K = 1$. In all the simulations we have used the Monte Carlo method with $K = 1$.

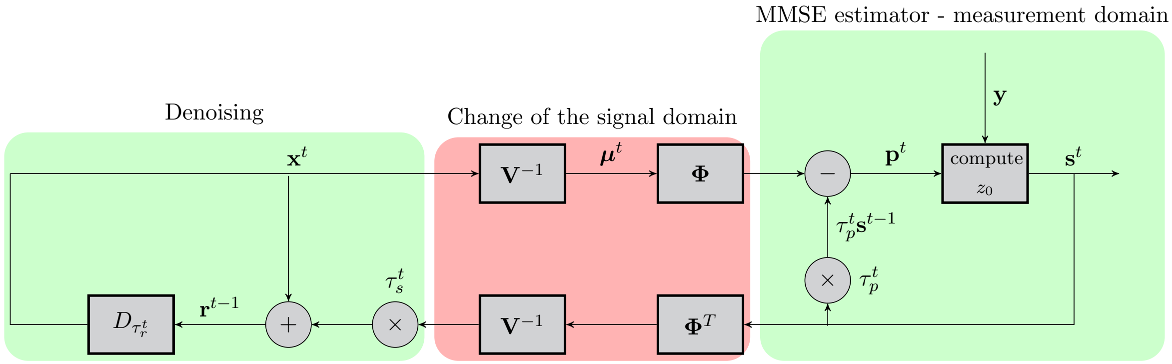



Figure 2: Block diagram of the D-GAMP-CT framework highlighting the 3 steps: 1) Denoising the signal estimate; 2) Preconditioning: change of the signal domain; 3) MMSE estimator for the non linear Poisson noise model.

With this method, the calculation of the Onsager term is more efficient since it requires only one more application of the denoiser. Moreover, it follows from Eq. (4.10) that the denoiser $D_{\tau_r^t}(\cdot)$, introduced in Section 4.3 acts on the high pass filtered image $\mathbf{x}$, whose expression is in Eq. (4.2).

The block diagram for the mean calculation of the proposed D-GAMP-CT algorithm is shown in Fig. 2; each iteration flow can be decomposed in 3 main steps: the MMSE estimation for the Poisson noise channel of the output vector $\mathbf{p}^t$, the preconditioning, which involves a change of the signal domain, and the denoising of the signal estimate. Fig. 2 graphically describes the steps for updating the mean vector variables of D-GAMP-CT algorithm listed in Algorithm 4.1; the denoising block corresponds to lines 15 and 18, while the application of the preconditioning matrix is needed in line 4 and 15 and finally the MMSE estimation in the measurement domain corresponds to lines 6, 9, 10.

**4.4. State evolution of D-GAMP-CT.** A significant characteristic of GAMP is that the MSE performance can be precisely predicted by a scalar SE analysis, with i.i.d. Gaussian random system matrices in the large system limit [49]; in particular, the GAMP SE formulation extends the AMP SE to arbitrary noise distributions.

In addition, if a generic denoiser is used within the AMP (D-AMP) iterations as in Eq. (4.10), it is shown heuristically in [35] that the MSE can be similarly predicted by the SE and, recently, a rigorous derivation of the SE for D-AMP is derived in [7].

The heuristic SE equations for the proposed D-GAMP-CT are based on the GAMP SE derivation [49] where the signal to estimate lies in the preconditioned domain and a denoiser is utilized as the non-linear input function. We should stress that a rigorous analysis for denoising GAMP has not yet been derived and that the SE analysis of D-AMP cannot be directly applied to denoising GAMP.

**Algorithm 4.1** D-GAMP-CT: Denoising Preconditioned Approximate Message Passing

1: **Initialization**: set $t = 0$, $\mathbf{r}^0 = \mathbf{0}$, $\mathbf{x}^0 = \mathbf{0}$, $\tau_x^0 = 1$
2: **for** $1, \ldots, T_{max}$ **do**
3: Step 1: Estimate in the measurement domain
4: $\mathbf{z}^t = \mathbf{\Phi}\mathbf{V}^{-1}\mathbf{x}^t$
5: $\tau_p^t = \frac{1}{M}\|\mathbf{\Phi}\mathbf{V}^{-1}\|_F^2\tau_x^t$
6: $\mathbf{p}^t = \mathbf{z}^t - \tau_p^t\mathbf{s}^{t-1}$
7:
8: Step 2: Poisson noise model
9: $\mathbf{z}_0^t = \mathbb{E}_{p(\mathbf{z}^t|\mathbf{p}^t,\mathbf{y},\tau_p^t)}[\mathbf{z}^t|\mathbf{p}^t, \mathbf{y}, \tau_p^t]$
10: $\mathbf{s}^t = \frac{\mathbf{z}_0^t - \mathbf{p}^t}{\tau_p^t}$
11: $\tau_s^t = \frac{1}{M\tau_p^t}\sum_{i=1}^{M}\left[1 - \frac{\mathrm{Var}(z_i^t|p_i^t, y_i, \tau_p^t)}{\tau_p^t}\right]$
12:

MMSE estimation in the measurement domain.

13: Step 3: Estimate in the signal domain
14: $\frac{1}{\tau_r^t} = \frac{1}{N}\|\mathbf{\Phi}\mathbf{V}^{-1}\|_F^2\tau_s^t$
15: $\mathbf{r}^t = \mathbf{x}^t + \tau_r^t\mathbf{V}^{-1}\mathbf{\Phi}^T\mathbf{s}^t$
16:
17: Step 4: Denoising step
18: $\mathbf{x}^{t+1} = D_{\tau_r^t}(\mathbf{r}^t)$
19: $\tau_x^{t+1} = \tau_r^t D'_{\tau_r^t}(\mathbf{r}^t)$

MAP/MSE estimation in the input domain.

20: **end for**
21: **return** $\boldsymbol{\mu}^t = \mathbf{V}^{-1}\mathbf{x}^t$

The D-GAMP-CT SE equations follow the GAMP SE formulation [49]; for the input and output vectors, we define 2 sets of vectors

$$\theta_r^t = (\mathbf{x}, \mathbf{r}^t, \mathbf{x}^t) \tag{4.15}$$

$$\theta_p^t = (\mathbf{z}, \mathbf{z}^t, \mathbf{y}, \mathbf{p}^t) \tag{4.16}$$

$\theta_p^t$ contains the components of the true and unknown vector $\mathbf{z}$, its D-GAMP-CT estimates $\mathbf{z}^t$ and $\mathbf{p}^t$ (line 6) and the observed measurement vector $\mathbf{y}$, while $\theta_p^t$ contains the unknown input vector $\mathbf{x}$ and the D-GAMP-CT estimates $\mathbf{x}^t$, and $\mathbf{r}^t$ (line 15). The main results in [49] state that for a fixed iteration $t$ and $N \to +\infty$ the joint empirical distribution of the elements for the vectors $\theta_r^t$ and $\theta_p^t$ converges empirically with second-order moments $PL(2)$ to the random vectors as

$$\lim_{N\to+\infty} \theta_r^t \overset{PL(2)}{=} \hat{\theta}_r^t(\xi_r^t) = (X, \hat{R}^t, \hat{X}^{t+1}) \tag{4.17}$$

$$\lim_{N\to+\infty} \theta_p^t \overset{PL(2)}{=} \hat{\theta}_p^t(\mathbf{K}_p^t) = (Z, \hat{Z}^t, Y, P^t) \tag{4.18}$$

$$\lim_{N\to+\infty} \tau_r^t = \hat{\tau}_r^t, \qquad \lim_{N\to+\infty} \tau_p^t = \hat{\tau}_p^t, \tag{4.19}$$

with

$$\hat{R}^t = X + V^t,\ V^t \sim \mathcal{N}(0, \xi_r^t) \quad \text{and} \quad \hat{X}^t = D_{\hat{\tau}_r}(\hat{R}^t) \tag{4.20}$$

for the input vector estimation and

$$(Z, P^t) \sim \mathcal{N}(0, \mathbf{K}_p^t),\ \hat{Z}^t = \mathbb{E}_{p(\mathbf{z}^t|\mathbf{p}^t,\mathbf{y},\hat{\tau}_p^t)}[P^t, Y, \hat{\tau}_p^t] \tag{4.21}$$

for the output vector. The SE equations in [49, Algorithm 3] produce a recursive scheme for calculating the parameters $\xi_r^t$, $\hat{\tau}_r$ of the distributions $\hat{\theta}_r^t$ and $\mathbf{K}_p^t$, $\hat{\tau}_p^t$ for $\hat{\theta}_p^t$. In the case $p_{Y|Z}$ matches the true distribution in Eq. (4.3), then it results that

$$\hat{\tau}_r^t = \xi_r^t = -\mathbb{E}^{-1}\left[\frac{\partial}{\partial p^t} g_{\text{out}}(P^t, Y, \hat{\tau}_p^t)\right] \tag{4.22}$$

where the expectation is taken over $\hat{\theta}_p^t(\mathbf{K}_p^t)$ with

$$\mathbf{K}_p^t = \begin{bmatrix} \hat{\tau}_x^0 & \hat{\tau}_x^0 - \hat{\tau}_p^t \\ \hat{\tau}_x^0 - \hat{\tau}_p^t & \hat{\tau}_x^0 - \hat{\tau}_p^t \end{bmatrix} \tag{4.23}$$

and $\hat{\tau}_x^0$ is set with an initial value and $\hat{\tau}_p^t = \beta\hat{\tau}_x^t$, $\quad \mathbf{K}_p^t = \beta\mathbf{K}_x^t$ with $\beta = \frac{M}{N}$. Following the derivation in [24, 49], the error and sensitivity functions are defined.

The error functions characterize the MSEs of the denoiser under Gaussian noise while the sensitivity functions describe the expected divergence of the estimator. The parameter $\hat{\tau}_x^t$ depends on both the error and sensitivity functions as follow. For the class of denoising functions $D_{\hat{\tau}_r}(\cdot)$ that are uniformly Lipshitz and convergent under Gaussian noise, which includes several non-separable denoisers [24], the sensitivity function is defined as

$$\hat{\tau}_x^t = \mathcal{A}_{in}(\hat{\tau}_r^t, \xi_r^t) = \lim_{N\to\infty} \langle \nabla D_{\hat{\tau}_r}(\mathbf{x} + \mathbf{v}^t) \rangle, \quad \mathbf{v}^t \sim \mathcal{N}(0, \xi_r^t \mathbf{I}) \tag{4.24}$$

and the error function is

$$\mathcal{E}_{in}(\hat{\tau}_r^t, \xi_r^t) = \lim_{N\to\infty} \frac{1}{N} \|D_{\hat{\tau}_r}(\mathbf{x} + \mathbf{v}^t) - \mathbf{x}\|^2, \quad \mathbf{v}^t \sim \mathcal{N}(0, \xi_r^t \mathbf{I}) \tag{4.25}$$

Unfortunately, the SE prediction is only valid in the random large system limit and therefore one may wonder what its relevance is in the considered CT problem. Here we argue that the empirical accuracy of the SE predictions provides an insight into the validity of D-GAMP-CT approximations when applied to such general linear models.

In particular, we claim that the small discrepancy is mainly due to the fast that the SE is derived under the "matched" condition while we calculate the posterior mean $\mathbf{z}_0^t$ and $\tau_s^t$ by the Laplace approximation in Eq. (4.6).

In Section 7, we present empirical evidence that the SE for D-GAMP-CT based on a real CT dataset provides an excellent prediction of the actual MSE achieved by D-GAMP-CT at each iteration.

## 5. Comparison with Other Methodologies.

**5.1. Vector AMP (VAMP).** Recently a new message passing algorithm VAMP (or its generalization VGAMP [60]) has been proposed which enjoys convergence guarantees for a larger class of random system matrices, i.e. right-orthogonal invariant. VAMP has been succesfully used in imaging application, like CT [56], and inverse scattering [61]. One difficulty within VAMP algorithm relies on the fact that its implementation requires either to compute the SVD of the system operator, or computing the covariance of the LMMSE estimator, i.e. inverting an high-dimensional symmetric matrix. Therefore, VAMP is particularly appealing for problems where it can be possible to compute the SVD of the system operator, either because it is available in matrix form, or because it can be decomposed by a fast orthogonal operator, like FFT, which leads to a fast computation of the trace of the inverse LMMSE covariance matrix. While in MRI it is possible to exploit the FFT form of the operator, unfortunately in CT, it is not generally possible to have a matrix form operator and therefore it becomes time consuming either computing the SVD off-line or estimating the inverse of the LMMSE covariance matrix.

**5.2. VAMP with Signal Whitening.** As described in the Introduction 1.2, different practical approaches have been proposed to handle non-random matrices within AMP or VAMP framework. An approach is the randomization of the input signal or scrambling its sample locations (or flipping its sample signs), then applying the sensing matrix on the randomized samples and finally, sub-sampling the resulting transform coefficients. Randomization methods de-correlate the signal with the sensing matrix, but generally this can be applied only with orthogonal sensing matrices, like Fourier matrix, but it does not apply to the Radon matrix for the reasons explained in Section 2.4. Furthermore, pre-randomizing the input might be physically not possible to implement and computationally inefficient.

A more efficient method presented in [58] is based on whitening the input signal and it is designed specifically for low-frequency Fourier matrix by using an appropriate wavelet transform as whitening operator. Instead, the aim of the proposed method based on preconditioning is to construct a new operator, close to be orthogonal, in an efficient manner, i.e. by exploiting the FFT. An interesting alternative to explore in the future is to use a wavelet transform $\mathbf{T}$ as whitening operator in conjunction with a Radon operator $\mathbf{A}$. Several wavelets, e.g., Haar, Daubechies, can achieve $O(N)$ complexity and the overall complexity of the composition $\mathbf{AT}$ would be $\mathcal{O}(N \log N)$. Similarly, the proposed Fourier-based preconditioning enjoys low $\mathcal{O}(N \log N)$ computational cost.

**5.3. Plug-and-Play Optimization.** Recent works [41, 30] have explored the Plug-and-play (PnP) incorporation of modern denoisers within the explicit regularization objective function in Eq. (2.9), as described in Section 2.3, or an alternative approach called regularization by denoising [55]. PnP approaches achieve state of the art recovery results in imaging applications even if they do not minimize an explicit MAP objective function [30] as for the regularization by denoising approach it is not completely understood what underlying objective function is being minimised by these algorithms - see the clarifications and new interpretations presented in [54]. We have compared our proposed framework with the Plug-and-Play ADMM (PnP-ADMM) optimization algorithm [63] in Sections 6 and 7.

**6. Simulation Results with Numerical Phantom.** We discuss the numerical results for 2D CT reconstruction using the "2016 NIH-AAPM-Mayo Clinic Low Dose CT Grand Challenge" full dataset as ground-truth; the slice 170 of dimension $(512 \times 512)$ from the "L067_full_1mm" acquisition is shown in fig. 3(b) and we have simulated a fan beam geometry, depicted in Fig. 3(a); we consider $\lfloor \frac{N}{5} \rfloor = 102$ views in the sinogram domain, obtained from a regular angular undersampling of the full projection measurements (1024 views $= 2N$), resulting in, approximately, 10 times undersampling ratio. The CT projection and back-projection operators are implemented using the ASTRA Toolbox [66].

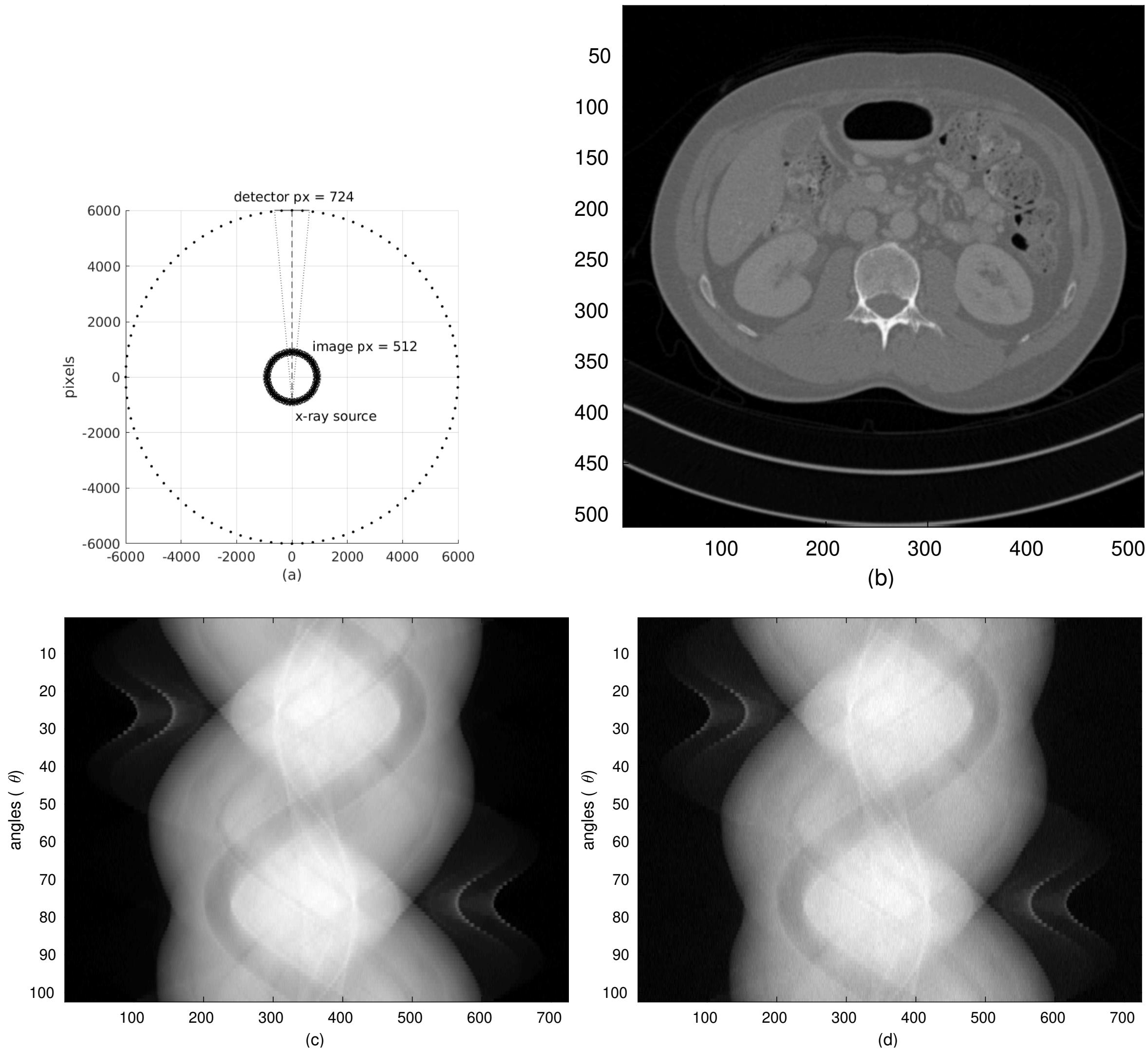



Figure 3: (a) Fan-beam geometry, (b) AAPM phantom - dataset L067_full_1mm, slice 170, (c) Sinogram for normal dose, $I_0 = 10^5$, (d) Sinogram for low dose, $I_0 = 10^4$.

The simulations include the Poisson noise model with different levels of intensity: an initial intensity of $I_0 = 10^5$, which is referred to as normal dose in the toolbox, and $I_0 = 10^4$ for the low dose case. The sparse views sinograms, for the 2 levels of intensity, are shown in Figs. 3(c)-(d) where it is worth noticing the low values in case of low dose; we will show that the Gaussian approximation of the CT noise is less effective with low beam intensity.

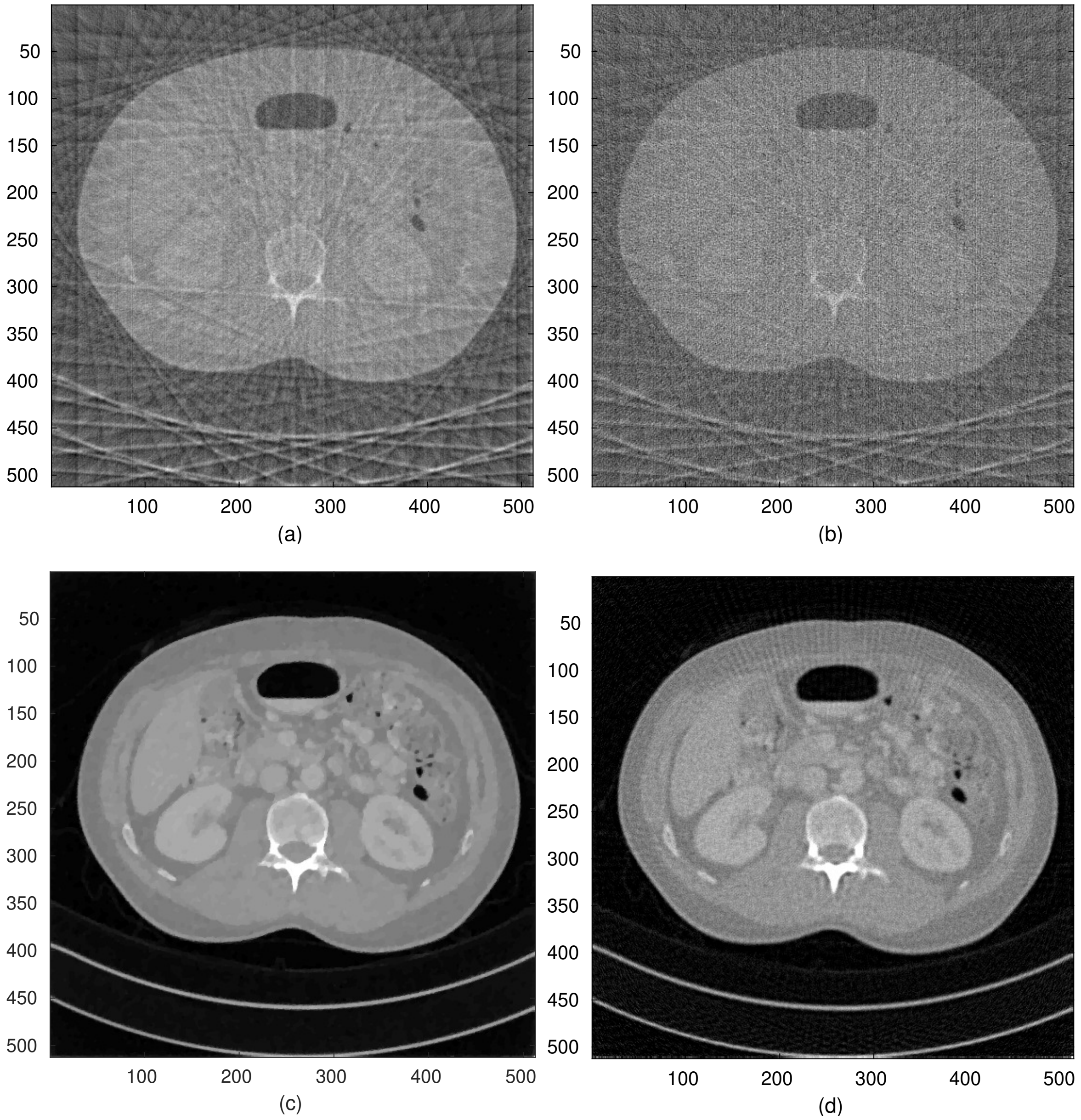



Figure 4: (a) FBP Normal dose, (b) FBP low dose. Normal dose: (c) BM3D-GAMP-CT, (d) BM3D-ADMM-WLS (PnP).

**6.1. Comparison D-GAMP-CT and Plug-and-play algorithms.** Figs. 4(a)-(b) show the FBP with ramp filter, for the normal and low photon intensities, which produces very poor reconstructions with strong streaking artifacts. In Figs. 4(c)-(d) are shown the reconstruction results for the normal dosage obtained using respectively BM3D-GAMP-CT algorithm 4.1, which reaches the convergence in 10 iterations, and the BM3D plug-and-play algorithm with WLS data fidelity term, implemented using the ADMM solver (BM3D-ADMM-WLS). The BM3D-GAMP-CT is built upon the GAMP Toolbox [21] and D-AMP Toolbox [20], while the BM3D denoiser is implemented using the Matlab toolbox [15] and the BM3D plug-and-play algorithm is implemented using [67] and is used as the reference reconstruction algorithm to compare with our proposed method. The image denoising algorithm BM3D is used as the denoiser in D-GAMP-CT since it provides good reconstruction performance and keeps computation time reasonable. It is worth noting that from Figs. 4(c)-(d), BM3D-GAMP-CT achieves a better qualitative reconstruction compared to BM3D-ADMM-WLS (PnP), whose output retains streaking noise artifacts in the inner region probably due to the the rays intercepting the hard tissue or bones. In Figs. 5(a)-(b) the results with low dose are shown for, respectively, BM3D-GAMP-CT and BM3D-ADMM-WLS. It is important to highlight that, in this case, the weighted Gaussian noise approximation, is not accurate due to the presence of zero values in the sinogram related in particular to the rays intercepting the bones. Taking the logarithm of the measurement leads to errors, especially in the region surrounded by hard tissue/bones; this is also confirmed quantitatively in Table 1. For a quantitative comparison, we have chosen the PSNR as the metric, defined as the ratio between the ground truth and the mean square error of the estimation, and the SSIM metric [71] which scores images in the interval $[0, 1]$, where a higher index represents better quality.

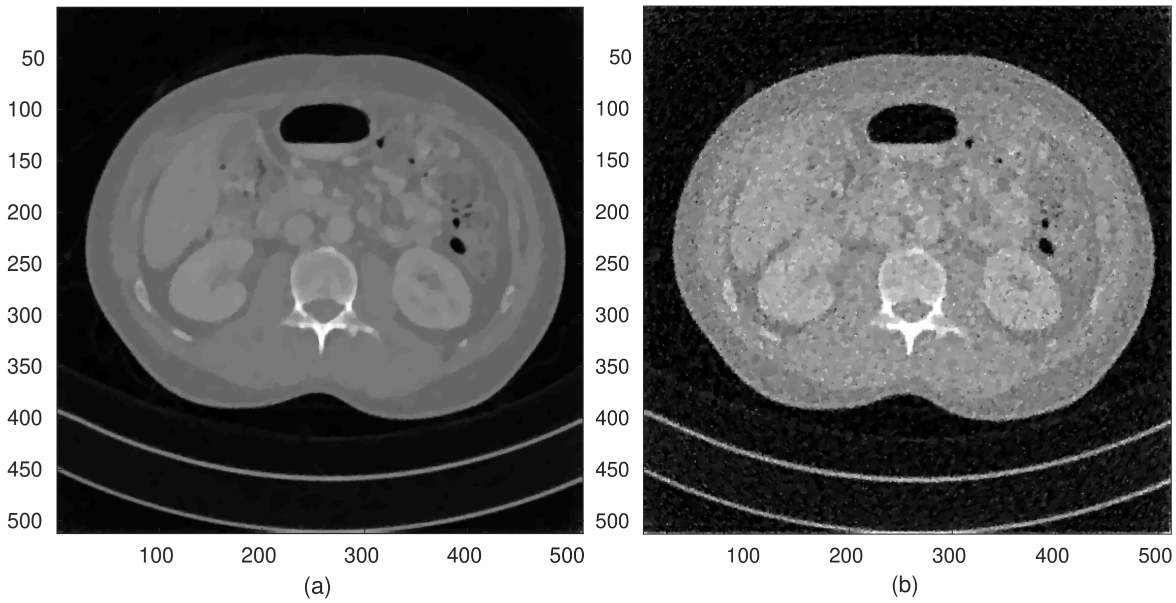



Figure 5: Low dose: (a) BM3D-GAMP-CT, (b) BM3D-ADMM-WLS (PnP).

Finally, it is worth comparing the proposed framework with a first order optimization solver for the regularized Poisson likelihood CT model. From the definition of the NLL function for a mixed Poisson Gaussian CT noise model in Eq. (2.6), we consider an approximation of Eq. (2.6) where we model the Poisson noise, neglecting the Gaussian noise ($\boldsymbol{\epsilon} \approx \mathbf{0}$); therefore, the associated NLL function can be rewritten in vector form as

$$L(\boldsymbol{\mu}) =< I_0 e^{-\boldsymbol{\Phi}\boldsymbol{\mu}}, \mathbf{1} > - < \mathbf{y}, \log(I_0 e^{-\boldsymbol{\Phi}\boldsymbol{\mu}}) > \tag{6.1}$$

The original regularized MAP Poisson objective function can be expressed as

$$\hat{\boldsymbol{\mu}}_{NLL} = \arg\min_{\boldsymbol{\mu}} L(\boldsymbol{\mu}) + R(\boldsymbol{\mu}) + \chi_B(\boldsymbol{\mu}) \tag{6.2}$$

where the non-negativity constraint on $\boldsymbol{\mu}$ is enforced by the characteristic function $\chi_B(\boldsymbol{\mu})$ on the set $B = \{\boldsymbol{\mu} : \mu_i \geq 0, \forall i\}$. We applied ADMM with splitting to solve (6.2) as it has been proposed in [16]. Figures 6(a)-(b) show respectively the results with normal and low dose by applying BM3D-ADMM-NLL (with BM3D denoiser, the plug-and-play ADMM minimizes a different MAP cost function compared to (6.2) with no explicit regularization $R(\mu)$).

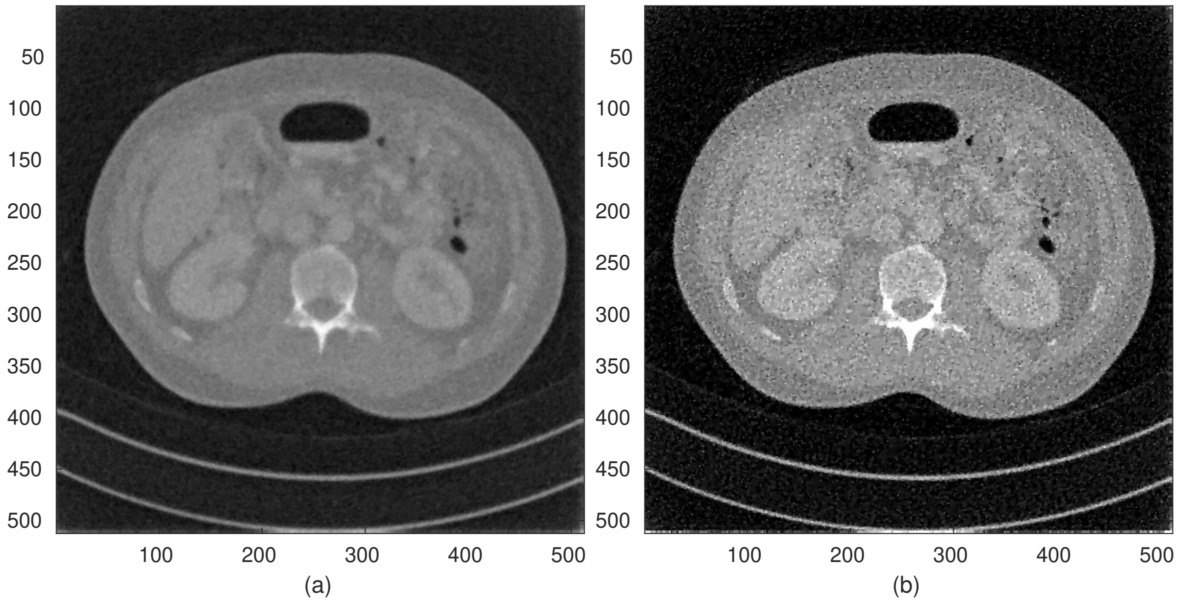



Figure 6: BM3D-ADMM-NLL (PnP): (a) Normal Dose, (b) Low Dose.

Table 1 also reports the quantitative results in terms of PSNR and SSIM for both low and high dose intensity values. The convergence plots in terms of MSE (dB) against iterations of the BM3D-GAMP-CT, BM3D-ADMM-WLS, BM3D-ADMM-NLL algorithms are shown in Fig. 7. In both normal and low dose scenario, it can be seen that BM3D-GAMP-CT produces a better quantitative reconstruction in terms of PSNR (dB) compared to BM3D-ADMM-WLS, and it requires lower total running time. For computational time evaluation, the simulations are run on an Intel®Xeon 2GHz machine using 8 cores.

Table 1: PSNR and time comparison

| Algorithms | PSNR [dB] | SSIM | Time |
|---|---|---|---|
| **Low photon intensity:** $I_0 = 10^4$ | | | |
| FBP | 31.5 | 0.36 | 45 sec |
| BM3D-ADMM-WLS (PnP) | 58.2 | 0.82 | 7.8 min |
| BM3D-ADMM-NLL (PnP) | 60.2 | 0.86 | 8 min |
| BM3D-GAMP-CT | 64.4 | 0.95 | 4.5 min |
| **High photon intensity:** $I_0 = 10^5$ | | | |
| FBP | 40.2 | 0.60 | 45 sec |
| BM3D-ADMM-WLS (PnP) | 65.6 | 0.88 | 7.8 min |
| BM3D-ADMM-NLL (PnP) | 67.1 | 0.91 | 8 min |
| BM3D-GAMP-CT | 70.4 | 0.97 | 4.5 min |

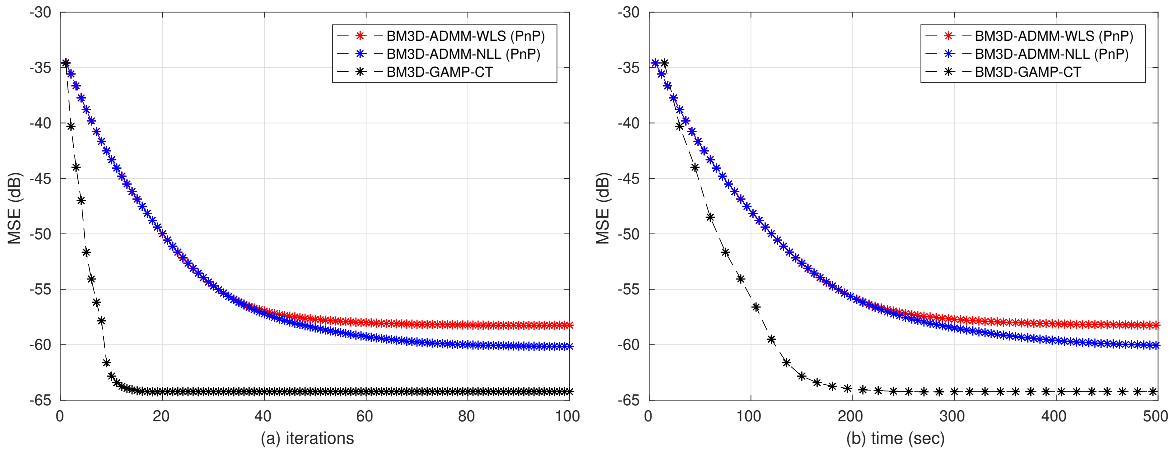



Figure 7: Comparison between BM3D-GAMP-CT, BM3D-ADMM-NLL and BM3D-ADMM-WLS: (a) MSE vs iterations; (b) MSE vs running time (sec).

**6.2. Comparison computational cost and running time.** From Algorithm 4.1 it is possible to estimate the order of complexity of the BM3D-GAMP-CT algorithm. At each iteration, 2 matrix vector multiplications are required in lines 5 and 14. For both operators $\widetilde{\mathbf{\Phi}}$ and $\widetilde{\mathbf{\Phi}}^T$ defined in Eq. (4.1), the matrix multiplication has a computational complexity of $\mathcal{O}(N \log N)$, since it is defined as a composition of element-wise operator of complexity $\mathcal{O}(N)$ and the operator $\mathbf{A}$ of complexity $\mathcal{O}(N \log N)$. The computation of the output vector mean $\mathbf{z}_0^t$ (line 9) and the variance $\tau_s^t$ does not involve matrix vector multiplications but summation of order $N$ as described in Appendix A. In addition, each iteration requires the application of one denoising function in (line 18) whose complexity depends on the actual implementation, since it is treated as a black box and the update of the variance $\tau_x^t$ requires to compute the divergence of the denoiser $D'_{\tau_\mathbf{r}}(\mathbf{r})$ which is implemented by Monte Carlo SURE [46] whose complexity is equivalent to one denoising function. On the other hand, BM3D-ADMM-WLS algorithm also requires at each iteration 2 matrix vector multiplications, solving one optimization subproblem with Conjugate Gradient (CG) and applying one denoiser (instead of complexity of 2 denoising functions required by BM3D-GAMP-CT). In our simulations shown in Fig. 7(b), the running time per iteration of BM3D-GAMP-CT is higher than the one of BM3D-ADMM-WLS because even the optimized implementation of BM3D [15] has an higher cost compared to the CG solver. However, the total running time of BM3D-GAMP-CT is lower (around 3 min as shown in Table 1 and Fig. 7(b)) because fewer iterations are needed to converge. Finally, compared to BM3D-ADMM-WLS the computation for the BM3D-ADMM-NLL solver with splitting requires to additionally inverting a circulant matrix by conjugate gradient whose complexity is at least of order $\mathcal{O}(N \log N)$.

**6.3. Proximal-based TV denoiser.** We show the reconstruction results using a different denoiser, the proximal Total Variation (TV) and we compare the result using a $\|\cdot\|_{(\mathbf{V}^T\mathbf{V})^{-1}}$ matrix norm instead of the $L_2$ norm for the proximal-based TV denoiser to properly handle the input signal in the preconditioned domain [9]. Fig. 8(b) shows the running time of $\text{prox}_{(\mathbf{V}^T\mathbf{V})}$TV-GAMP-CT, proxTV-GAMP-CT and proxTV-ADMM-WLS; we observe that the total running time of the proxTV-GAMP-CT based algorithm is lower of the proxTV-PnP Plug-and-play and, together with the previous results with BM3D, D-GAMP-CT leads to a reduction in both number of iterations and total running time compared to the PnP approach, irrespective to the type of denoiser used. Furthermore, by comparing with the plot in Fig. 7(b) while the MSE accuracy at convergence achieved with proxTV is about 9dB worse compared to BM3D, the iteration time with proxTV is lower since the complexity of proxTV is only of order of $\mathcal{O}(N)$.

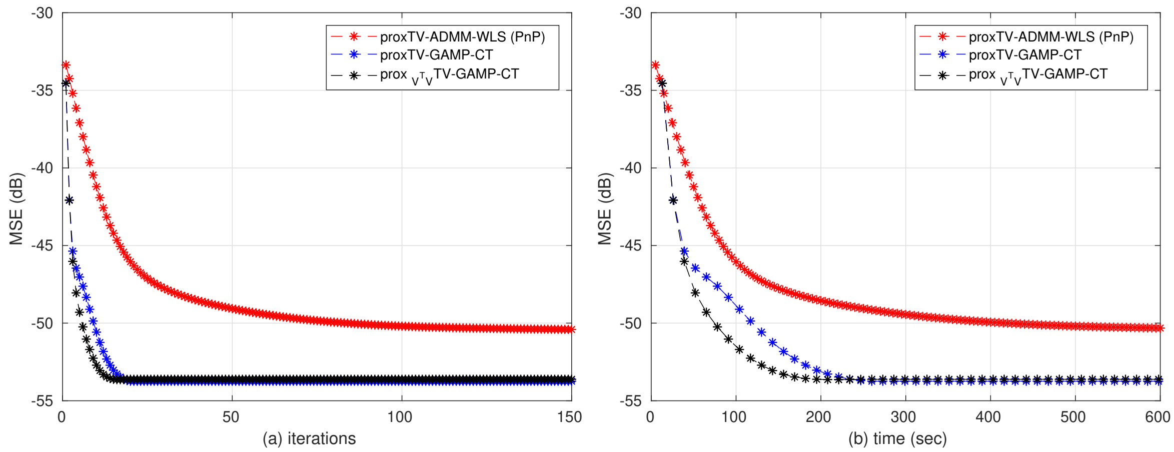



Figure 8: Comparison between $\text{prox}_{(\mathbf{V}^T\mathbf{V})}$TV-GAMP-CT, proxTV-GAMP-CT and proxTV-ADMM-WLS: (a) MSE vs iterations; (b) MSE vs running time (sec).

Interestingly, the $\text{prox}_{(\mathbf{V}^T\mathbf{V})}$TV-GAMP-CT, compared to the proxTV-GAMP-CT, leads to an improvement in the accuracy error only at earlier iterations, before convergence, while both converges close to the same MSE value. This is intuitively expected because the $\text{prox}_{(\mathbf{V}^T\mathbf{V})}$TV tends to better reduce the noise variance at earlier iterations since it takes into account the appropriate norm for the preconditioned space but at convergence the achieved accuracy is almost equivalent to the $l_2$ norm-based proxTV.

**6.4. Comparison between BM3D-GAMP-CT and BM3D-VAMP-CT.** We analyze the performance of the BM3D-GAMP-CT and BM3D-VAMP-CT which is an alternative type of message passing algorithm (as described in Section 5.1). The SVD-based implementation of VAMP requires the calculation of the eigenvalues of the symmetric matrix $\mathbf{A}\mathbf{A}^T$, with $\mathbf{A}$ being the forward system matrix. In the ASTRA toolbox, $\mathbf{A}$ is defined in form of an operator, for fast computations. Therefore, it is not efficient to estimate the eigenvalues in high dimensions since it requires to generate the $M$ columns of $\mathbf{A}\mathbf{A}^T$ by calculating $M$ times $A(A^T(\mathbf{e}_i))$, being $\mathbf{e}_i$ the $i$-th unity vector, $i = 1, \ldots, M$, and then calculating the eigenvalues By using the Matlab command eig, it takes around 10 min to generate the eigenvalues. We report an estimated value for the condition number $\kappa = \frac{\lambda_{max}}{\lambda_{min}} \sim 5 \times 10^4$. Moreover, Fig. 9 shows that both algorithms behave similarly and they converge at similar values of MSE (0.8 dB difference). This simulations seem not to be in disagreement with earlier results in [52](Figs. 3 and 4) where VAMP and GAMP with damping exhibit similar MSE behavior. From the simulation, both in terms of computation and accuracy, BM3D-GAMP-CT (with preconditioning) outperforms BM3D-VAMP-CT for this CT dataset.

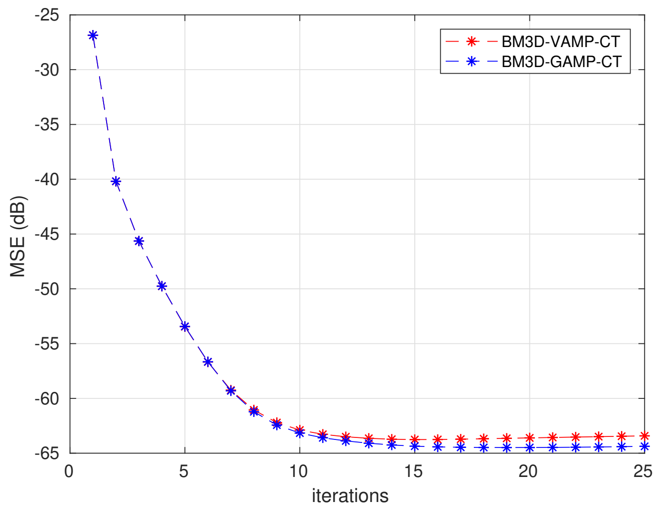



Figure 9: MSE plot: comparison between BM3D-VAMP-CT and BM3D-GAMP-CT.

**7. Experimental Results.** In this section, we investigate the reconstruction quality of D-GAMP-CT on real CT data. The D-GAMP-CT framework has been applied for CT reconstruction on real luggage scans obtained using Morpho CTX5500 Air Cargo dual energy system with fan beam CT geometry. This is a single-row scanner with 476 detector channels and a 80 cm field of view. For each transversal location, two slices were acquired one at 100 KVp, the other at 198 KVp; at each energy, the full acquisition of a single slice contains 720 views/projections. The reconstruction has been performed for each energy independently and here we consider only the results obtained for 100 kVp. The reconstructed images are of $512 \times 512$ array size. The total dataset contains 44 scanned slices (images).

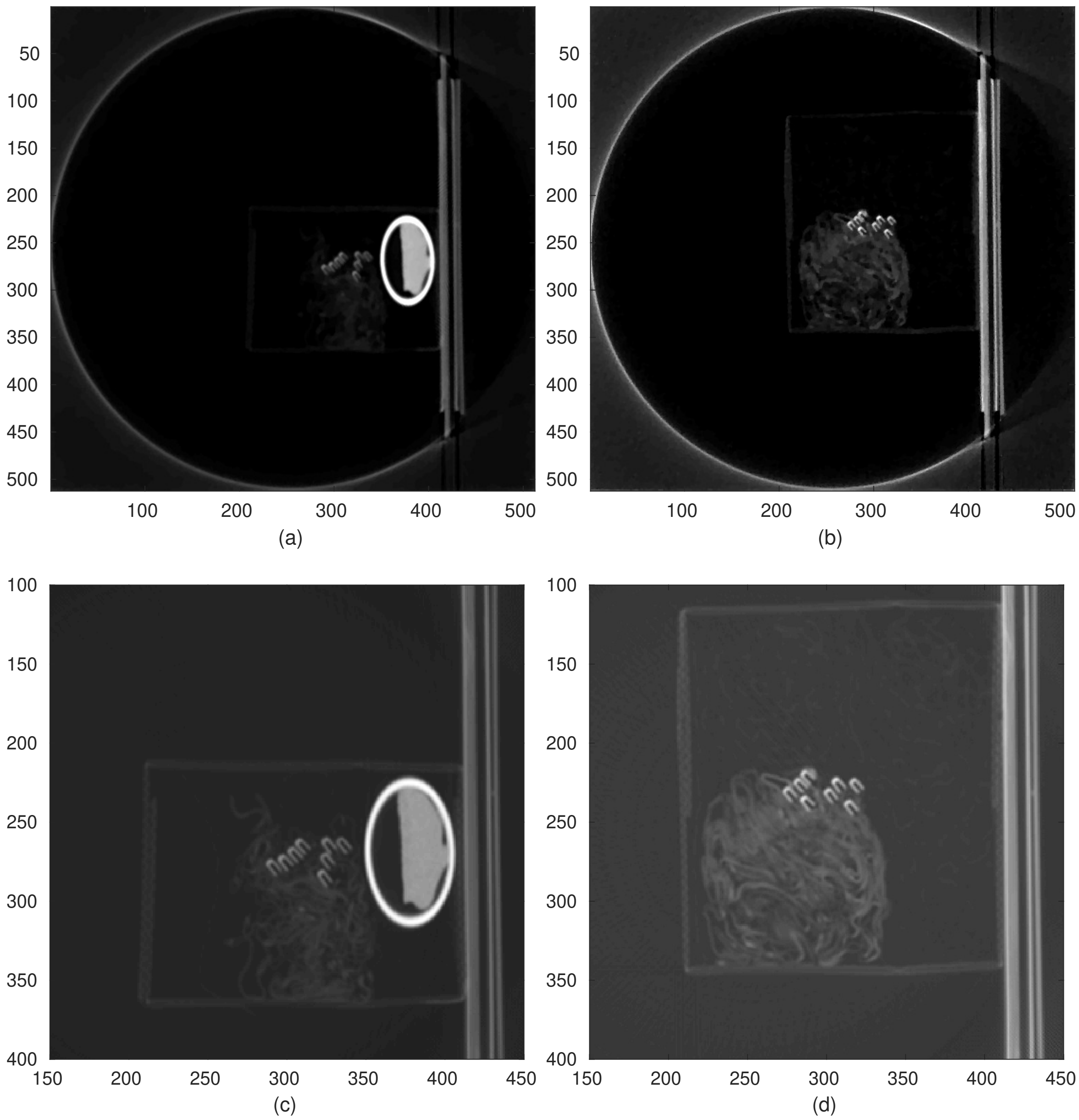



Figure 10: Full CT reconstruction and ROI using BM3D-GAMP-CT with preconditioning and Poisson noise model: (a)-(c) slice 10, (b)-(d) slice 42.

The results in Fig. 10 show two slices and the zoom around the region of interest (ROI) from the reconstruction with BM3D-GAMP-CT using 72 views regularly undersampled out of the full set of views constituted of 720 views. For computing the quantitative performances in terms of PSNR and SSIM, the FBP with full number of projections (720 views) is considered as a proxy for the ground truth. In the figure it is possible to see that the scanned object contains highly resolved metal staples, bottle of fluid, wires. The number of iterations for BM3D-GAMP-CT algorithms tends to converge in around 15 iterations as shown in Section 7.3. In Figure 11 we show the reconstruction for 2 slices of the entire volume (10 and 42) obtained with FBP using 72 views for one of the image slices.

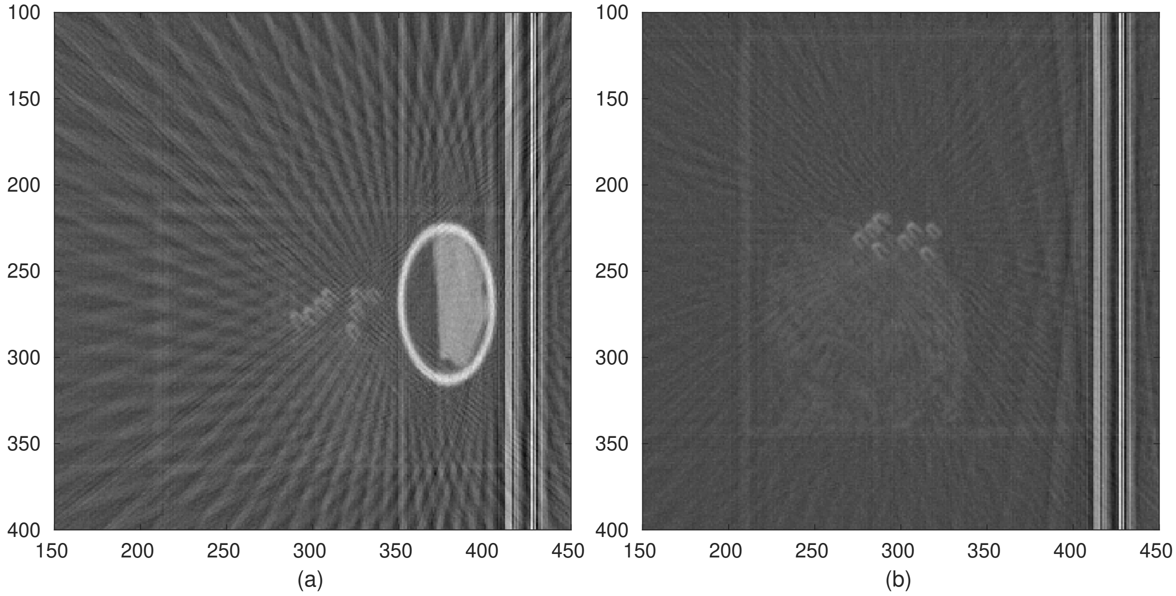



Figure 11: Filtered Back Projection with 72 projections: (a) slice 10; (b) slice 42.

**7.1. Role of the Onsager term.** Given the similarity between iterative shrinkage algorithms and the GAMP family of algorithms, it is interesting to evaluate the importance of the Onsager term $\tau_p^t \mathbf{s}^{t-1}$ in the D-GAMP-CT algorithm 4.1 to check whether it improves the reconstruction. Without the Onsager term the GAMP algorithm behaves like a denoising iterative thresholding algorithm [35]. The reconstruction for slices 10 and 42 without the Onsager term is shown in Fig. 12 which highlights a substantial reduction in performance in both cases, as it is also quantitatively confirmed by the PSNR value in table 2.

The Onsager term yields a PSNR improvement of 6 dB, for this particular CT reconstruction instance. Furthermore if we consider the algorithm with Onsager term, although the time per iteration is almost doubled because of the computation of an additional denoiser, the total running time is lower compared to the version without Onsager term because it converges in few iterations (15 respect to 40).

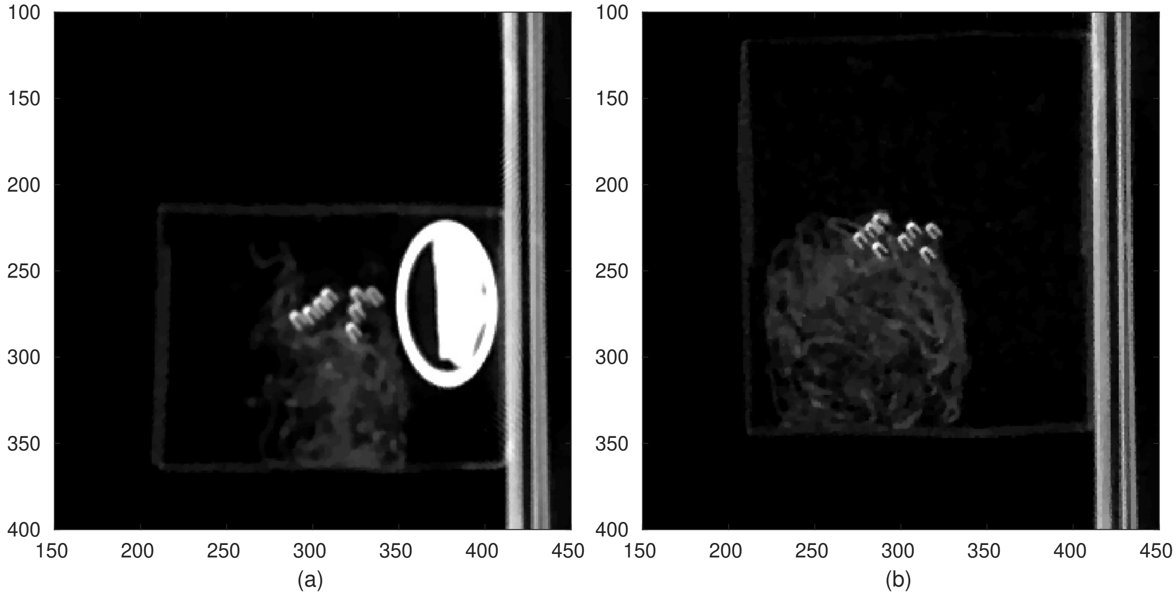



Figure 12: CT image reconstruction using BM3D-AMP without Onsager term for: (a) slice 10; (b) slice 42.

Table 2: PSNR for of BM3D-GAMP-CT with/without Onsager term

| **Algorithms** | **PSNR** [dB] | **SSIM** | **Time** |
|---|---|---|---|
| BM3D-GAMP-CT | 61.2 | 0.85 | 3.5 min |
| BM3D-GAMP-CT without Onsager term | 55.1 | 0.71 | 4 min |
| BM3D-GAMP with damping | 50.3 | 0.65 | 4.5 min |

**7.2. Role of the Preconditioner.** By applying the Denoising Generalized Approximate Message Passing algorithm, which includes the exponential Poisson noise model, the iteration estimates diverge. Therefore, for comparison we applied damping [59, 68] in estimating vectorised image at iteration $t$, i.e. $\mathbf{x}_t = \eta_x \mathbf{x}^t + (1-\eta_x)\mathbf{x}^{t-1}$, with $0 < \eta_x < 1$ the damping factor and on the estimated linear measurement vector, i.e. $\mathbf{s}^t = \eta_r \mathbf{s}^t + (1-\eta_r)\mathbf{s}^{t-1}$, with $0 < \eta_r < 1$, where the vector variables $\mathbf{x}$ and $\mathbf{s}$ are calculated as in Algorithm 4.1. In this case, the BM3D-GAMP-CT algorithm (without preconditioning) starts to converge for $\eta_r = 0.95$ and with a good amount of damping on the estimated $\mathbf{x}^t$, $\eta_x = 0.65$. Fig. 13 shows the reconstruction of damped BM3D-GAMP-CT and Table 2 reports the quantitative metrics. BM3D-GAMP-CT exhibits improved PSNR and SSIM over both BM3D-GAMP-CT without the Onsager term or with damping. Furthermore, the computational time both without Onsager term or with damping is higher since the algorithms require more iterations, 40 and 25 respectively, to converge compared to 15 iterations needed by BM3D-GAMP-CT. Regarding the behavior of the damping for ill-conditioned matrices, our results are coherent with previous simulations reported in [68](Fig. 1D) where the adaptive damping in GAMP does not prevent divergence for high values of condition number of the system matrix $A$. Therefore, damping is effective for low condition number (less than 10), while our Radon transform has condition number of order $\sim 10^4$. Therefore, our results confirm that damping is not very effective for high ill-conditioned matrix.

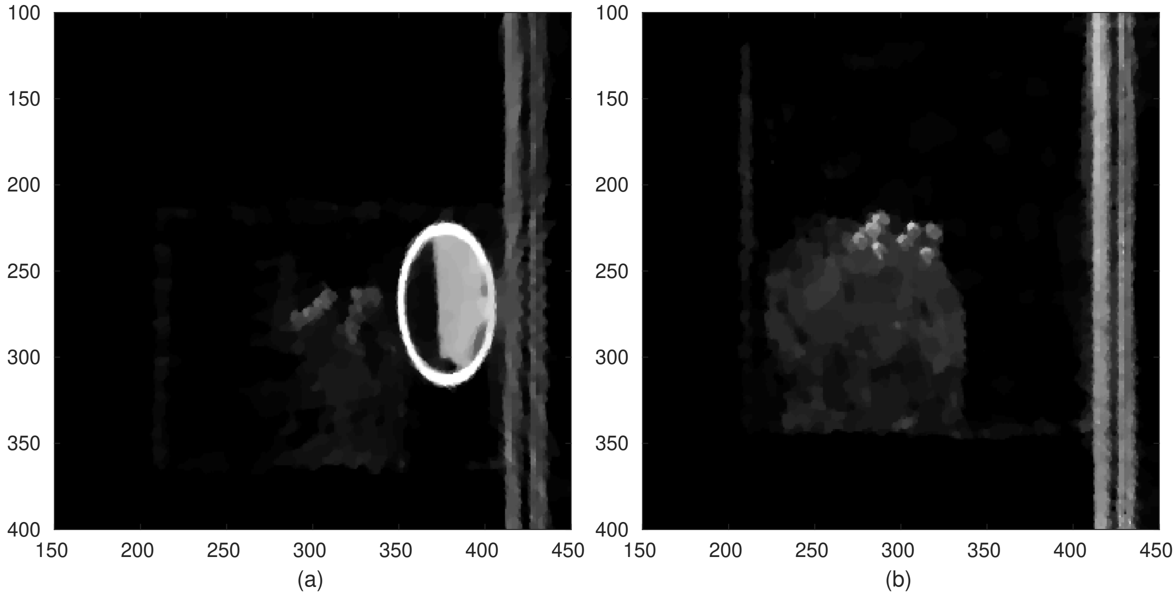
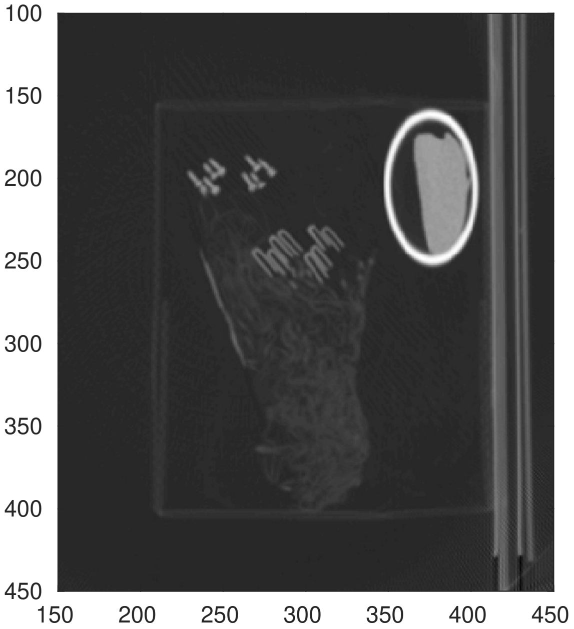


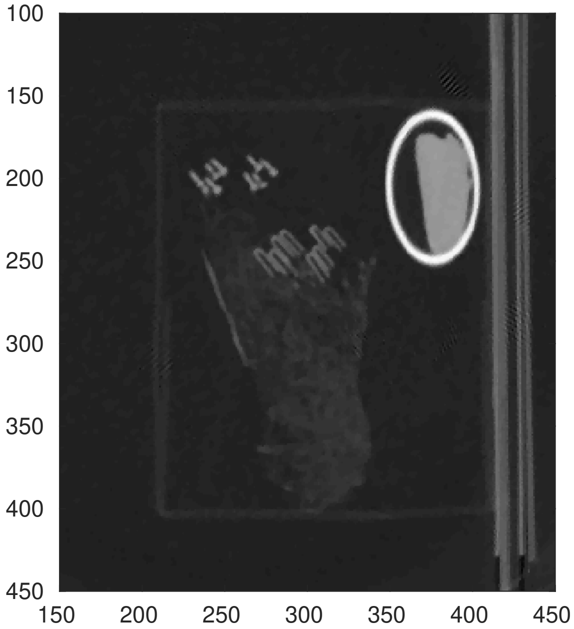

Figure 13: BM3D-GAMP-CT reconstruction with damping $\eta_x = 0.65$, $\eta_z = 0.9$ for: (a) slice 10; (b) slice 42.

**7.3. Comparison against Plug-and-play approach.** We present a comparison between D-GAMP-CT and PnP, using ADMM for solving the NLL minimization (6.2), using BM3D and TV denoisers. We are comparing 3 different methods: BM3D-GAMP-CT and the Plug-and-play approach with NLL data fidelity term BM3D-ADMM-NLL and TV-ADMM-NLL. Since the denoisers implicitly imposed in the cost function that BM3D-ADMM-NLL minimizes is different from the one of TV-ADMM-NLL, the 2 solvers yield different accuracies.

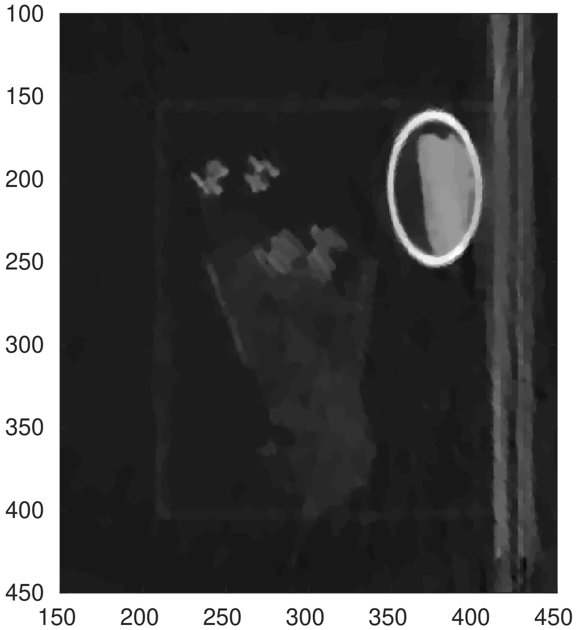
In Fig. 14, we show the qualitative results of one reconstructed slice. From inspection, it is clear that the reconstruction with BM3D-GAMP-CT better retains the details in the ROI. This is confirmed quantitatively in Figs. 15(a)-(b) which show the convergence of the MSE against iterations and running time. Table 3 contains the quantitative details in terms of PSNR, SSIM and time. If we analyze the algorithms using the same denoiser (BM3D), Fig. 15(a) shows that the BM3D-GAMP-CT (line in black) converges faster both in terms of number of iterations and total running time compared to BM3D-ADMM-NLL for Plug-and-Play optimization. Furthermore, at convergence BM3D-GAMP-CT yields to a reduction in PSNR of 3 dB.



(a)



(b)



(c)

Figure 14: Recontruction of slice 20 with: (a) BM3D-GAMP-CT, (b) BM3D-ADMM-NLL (PnP) (c) TV-ADMM-NLL (PnP) - NLL objective (6.2) with TV denoiser.

Table 3: PSNR of BM3D-GAMP-CT, BM3D-ADMM-NLL and TV-ADMM-NLL

| **Algorithms** | **PSNR** [dB] | **SSIM** | **Time** |
|---|---|---|---|
| BM3D-GAMP-CT | 61.3 | 0.85 | 3.5 min |
| BM3D-ADMM-NLL (PnP) | 58.2 | 0.78 | 11 min |
| TV-ADMM-NLL (PnP) | 56.8 | 0.72 | 6.5 min |

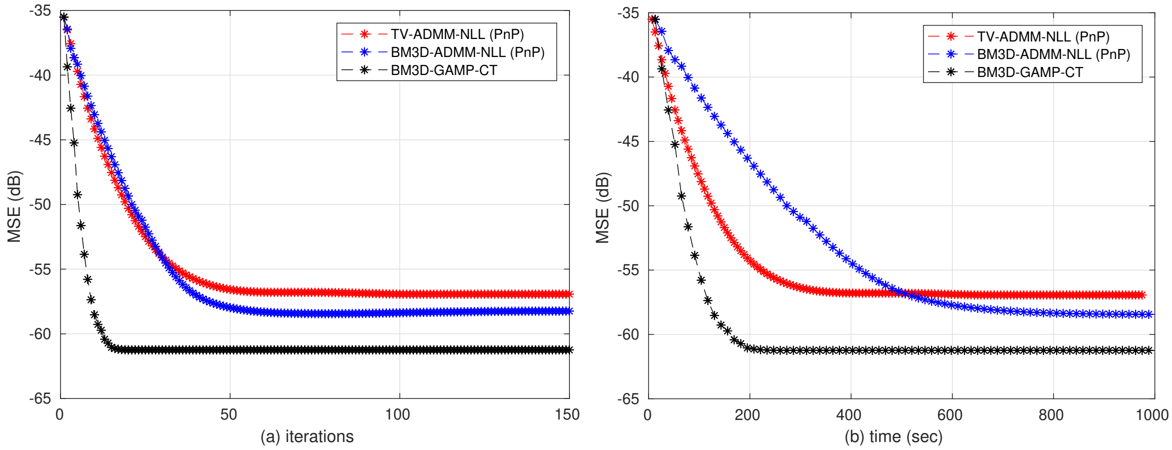



Figure 15: Comparison between BM3D-GAMP-CT, BM3D-ADMM-NLL and TV-ADMM-NLL: (a) MSE vs iterations; (b) MSE vs running time (sec).

We have tried a different denoiser, TV, for the PP framework and we can observe from Figs. 15 that at earlier iterations BM3D-ADMM-NLL and TV-ADMM-NLL perform similarly because the input estimate is highly noisy but at convergence the BM3D reaches higher PSNR and SSIM. Interestingly, the running time of TV-ADMM-NLL is lower compared to BM3D-ADMM-NLL since they converge almost at the same number of iterations but the TV denoiser is computationally faster than BM3D. However, BM3D-GAMP-CT is faster in total running time than TV-ADMM-NLL because BM3D-GAMP-CT converges in almost $\frac{1}{6}$ the number of iterations of TV-ADMM-NLL.

**7.4. State Evolution Analysis.** An important aspect of D-GAMP-CT is its internal variance estimate within the algorithm and in the SE equations. This not only provides an estimate of uncertainty with the algorithm, it also essentially allows the algorithm to adapt its "step size" on the fly [51]. It is therefore instructive to see how precise such an estimate is. If accurate, this term should ensure a fast convergence rate of the algorithm. Given the actual MSE estimate, taking as reference the full views FBP reconstruction, we can calculate the predicted MSE at the next iteration and compare with the actual estimate.

Our empirical results run on the experimental data acquired using the 2D fan beam CT geometry show that the SE prediction gives an accurate estimate of the true MSE of BM3D-

GAMP-CT at each iteration as depicted in Figure 16.

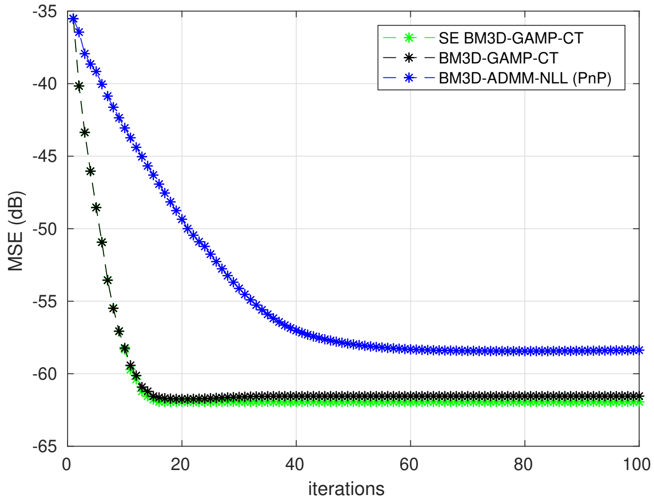



Figure 16: Deterministic state evolution and MSE estimates using BM3D-AMP-CT and BM3D-ADMM-NLL.

Fig. 16 shows the theoretical SE (prediction of the MSE) of the proposed D-GAMP-CT algorithm (in red) and the actual performance of the algorithm with BM3D. The 2 curves are almost matched; the little discrepancy is due to the fact that the SE is derived under the "matched" case indicated in Eq. (4.22), i.e. exact estimation of the variance of the MMSE for the Poisson noise while, as described in Eq. (4.6), we estimate it by using the Laplace approximation method. Moreover, this plot highlights that the Laplace approximation is accurate since the error with the SE is only around 0.1 dB. Furthermore, the Plug-and-play solver BM3D-ADMM-NLL (black curve) performs always worse than D-GAMP-CT and also the MSE behavior of BM3D-ADMM-NLL cannot be predicted by SE equations.

**8. Conclusions and Future Research.** In this work, we have presented a Generalized Approximate Message Passing type of iterative algorithm for solving X-ray CT reconstruction from a limited number of projections. The proposed framework relies on the design of an appropriate preconditioner for the ill-conditioned measurement matrix and a statistical model for the non linear Poisson measurement noise. In addition, exploiting the flexibility of the GAMP framework we can decouple the action of the preconditioner from the noise model, which is not possible with optimization solvers for minimizing the Plug-and-play PP-WLS objective function.

We have experimentally shown the important role of the Onsager term regarding reconstruction performance improvement and the ability of the state evolution analysis to estimate the current MSE through the iterations. Numerical results on simulations and experimental Cargo dataset demonstrate how the D-GAMP-CT framework provides high reconstruction accuracy and reduced running time compared to state-of-the-art Plug-and-play optimization iterative algorithms for CT reconstruction. In addition D-GAMP-CT allows different prior image models to be used on the signal by employing different denoisers.

Finally, further acceleration of the D-GAMP-CT may be possible utilizing the Ordered Subsets principle [19], however its implementation is not straight forward within the GAMP framework and is also left for future research.

**Acknowledgment.** The authors are grateful to Jonathan H. Mason from The University of Edinburgh for providing invaluable comments and for improving the quality of this work with insightful discussions. The authors thank Walter Garms, Morpho Detection Inc., for providing real CT datasets. This work was supported by U.S. Department of Homeland Security, Science and Technology Directorate, Explosives Division, BAA 13-05, Contract # HSHQDC-14-C-B0048, EPSRC platform grant EP/J015180/1 and the ERC Advanced grant, project C-SENSE, (ERC-ADG-2015-694888). The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 713683 (COFUNDfellowsDTU).

## Appendix A. Laplace method for approximating the posterior mean of the Nonlinear noise distribution.

In order to evaluate the expression in (4.6), we write the ration of integrals in the following form (where we have not indicated the iteration $t$ for notation simplicity)

$$\mathbb{E}(z_a|p_a, y_a, \tau_p) = \frac{\int_{\mathbb{R}_{\geq 0}} g(z_a) e^{\log p(y_a|z_a)} \pi(z_a) dz_a}{\int_{\mathbb{R}_{\geq 0}} e^{\log p(y_a|z_a)} \pi(z_a) dz_a} \tag{A.1}$$

where $\pi(z_a) = e^{-\frac{1}{2\tau_p}(z_a - \hat{p_a})^2}$ and $g(z_a) = z_a$. We set

$$L = \log \pi + \frac{1}{M} \log p(y_a|z_a) = -\frac{1}{2\tau_p}(z_a - p_a)^2 + \frac{1}{M}\left[ - z_a y_a - e^{-z_a} - \log(y_a!)\right]$$

$$\begin{aligned} L^* &= \log z_a + L = \log g(z_a) + \log \pi(z_a) + \frac{1}{M} \log p(y_a|z_a) \\ &= \log z_a - \frac{1}{2\tau_p}(z_a - p_a)^2 + \frac{1}{M}\left[ - z_a y_a - e^{-z_a} - \log(y_a!)\right] \end{aligned} \tag{A.2}$$

Therefore, the MMSE can be written as

$$\mathbb{E}(z_a|p_a, y_a, \tau_p) = \frac{\int_{\mathbb{R}_{\geq 0}} e^{M \cdot L^*} dz_a}{\int_{\mathbb{R}_{\geq 0}} e^{M \cdot L} dz_a} \tag{A.3}$$

We consider the probability density function $L(z_a)$ which we expect to have a peak at the point $z_{0_a}$ and the Taylor-expansion of $L(z_a)$ at $z_{0_a}$ is

$$L(z_a) \approx L(z_{0_a}) - \frac{1}{2}\frac{\partial^2 L(z_a)}{\partial z_a^2}(z_a - z_{0_a})^2 + \dots \tag{A.4}$$

The Laplace's method [65] is a way to approximate $L(z_a)$ by an unnormalized Gaussian and approximate the partition function $Z_P = \int L(z_a) dz_a$ with the one of the Gaussian

$$Z_Q = L(z_{0_a})\sqrt{2\pi} \tag{A.5}$$

The Laplace approximation leads to

$$\int e^{mL(z_a)}dz_a \approx \int e^{mL(z_{0_a})-m(z_a-z_{0_a})^2/(2\sigma^2)}dz_a$$

$$= \sqrt{2\pi}\sigma n^{-1/2}e^{mL(z_{0_a})} \tag{A.6}$$

with $\sigma^2 = -\frac{1}{L''(z_{0_a})}$; this integral form is similar to the one in Eq. (4.6) for the numerator and denominator respectively. Considering the denominator, we need to calculate the points where the derivative is zero in order to find $z_{0_a}$ :

$$\frac{\partial L(z_a)}{\partial z_a} = -\frac{1}{\tau_p}(z_a - p_a) - y_a + e^{-z_a}$$

with $y_a \in \mathbb{Z}_+$, $z_a = \left[\widetilde{\boldsymbol{\Phi}}\mathbf{x}\right]_a \in \mathbb{R}_{\geq 0}$; then, to find $\frac{\partial L(z_a)}{\partial z_a} = 0$, it results

$$-\frac{(z_a - p_a)}{\tau_p} - y_a + e^{-z_a} = 0$$

$$\log\left[-\frac{(z_a - p_a)}{\tau_p} - y_a\right] = z_a$$

Finally, the second derivative is

$$\left.\frac{\partial^2 L(z_a)}{\partial z_a^2}\right|_{z_{a_0}} = -\frac{z_{a_0}}{\tau^p} - e^{-z_{a_0}}$$

Similar procedure for the numerator ($\sigma^*$ and $z^*_{a_0}$); therefore, taking the ratio of the 2 approximations it yields to

$$\mathbb{E}[z_a|p_a] = \frac{\sigma^*}{\sigma}e^{L^*(z^*_{a_0})-L(z_{a_0})} \tag{A.7}$$

For the variance, given the approximation A.7, we can use the standard formula

$$\mathrm{Var}[z_a|p_a] = \mathbb{E}[z_a^2|p_a] - \mathbb{E}[z_a|p_a]^2 \tag{A.8}$$

## REFERENCES


[1] A. Averbuch, R. Coifman, D. Donoho, M. Israeli, and Y. Shkolnisky, *A framework for discrete integral transformations I - the pseudopolar fourier transform*, SIAM Journal on Scientific Computing, 30 (2008), pp. 764–784.

[2] A. Averbuch, R. Coifman, D. Donoho, M. Israeli, and J. Walden, *Fast slant stack: A notion of radon transform for data in a cartesian grid which is rapidly computible, algebraically exact, geometrically faithful and invertible*, SIAM Scientific Computing, (2001).

[3] A. Averbuch, R. Coifman, D. L. Donoho, M. Israeli, Y. Shkolnisky, and I. Sedelnikov, *A framework for discrete integral transformations II - the 2D discrete radon transform*, SIAM Journal on Scientific Computing, 30 (2008), pp. 785–803.

[4] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Transactions on Information Theory, 57 (2011), pp. 764–785.
[5] M. Bayati and A. Montanari, *The lasso risk for gaussian matrices*, IEEE Transactions on Information Theory, 58 (2011), pp. 1997–2017.
[6] M. Beister, D. Kolditz, and W. A. Kalender, *Iterative reconstruction methods in X-ray CT*, Physica medica, 28 (2012), pp. 94–108.
[7] R. Berthier, A. Montanari, and P.-M. Nguyen, *State evolution for approximate message passing with non-separable functions*, Information and Inference: A Journal of the IMA, 9 (2020), pp. 33–79.
[8] P. Bruyant, J. Sau, and J. Mallet, *Streak artifact reduction in filtered backprojection using a level line–based interpolation method*, Journal of Nuclear Medicine, 41 (2000), pp. 1913–1919.
[9] G. T. Buzzard, S. H. Chan, S. Sreehari, and C. A. Bouman, *Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium*, SIAM Journal on Imaging Sciences, 11 (2018), pp. 2001–2020.
[10] B. Cakmak, O. Winther, and B. H. Fleury, *S-AMP: Approximate message passing for general matrix ensembles*, in IEEE Information Theory Workshop (ITW), 2014, pp. 192–196.
[11] L. Calatroni and A. Chambolle, *Backtracking strategies for accelerated descent methods with smooth composite objectives*, arXiv preprint arXiv:1709.09004, (2017).
[12] S. Y. Chun, Y. Dewaraja, and J. A. Fessler, *Alternating direction method of multiplier for tomography with nonlocal regularizers*, IEEE Transactions on Medical Imaging, 33 (2014), pp. 1960–1968.
[13] N. H. Clinthorne, T. S. Pan, P. C. Chiao, W. L. Rogers, and J. A. Stamos, *Preconditioning methods for improved convergence rates in iterative reconstructions*, IEEE Transactions on Medical Imaging, 12 (1993), pp. 78–83.
[14] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising by sparse 3-d transform-domain collaborative filtering*, IEEE Transactions on Image Processing, 16 (2007), pp. 2080–2095.
[15] A. Danielyan, V. Katkovnik, and K. Egiazarian, *BM3D frames and variational image deblurring*, IEEE Transactions on Image Processing, 21 (2012), pp. 1715–1728.
[16] Q. Ding, Y. Long, X. Zhang, and J. A. Fessler, *Statistical image reconstruction using mixed poisson-gaussian noise model for x-ray ct*, arXiv preprint arXiv:1801.09533, (2018).
[17] D. Donoho, A. Maleki, and A. Montanari, *Message-passing algorithms for compressed sensing*, Proceedings of the National Academy of Sciences, 106 (2009), pp. 18914–18919.
[18] I. A. Elbakri and J. A. Fessler, *Statistical image reconstruction for polyenergetic X-ray computed tomography*, IEEE Transactions on Medical Imaging, 21 (2002), pp. 89–99.
[19] H. Erdogan and J. A. Fessler, *Ordered subsets algorithms for transmission tomography*, Physics in medicine and biology, 44 (1999), p. 2835.
[20] C. M. et al., *D-AMP toolbox*. https://github.com/ricedsp/D-AMP_Toolbox, 2017.
[21] S. R. et al., *Gampmatlab toolbox*. http://gampmatlab.wikia.com/wiki/Generalized_Approximate_Message_Passing, 2017.
[22] J. A. Fessler and S. Booth, *Conjugate-gradient preconditioning methods for shift-variant pet image reconstruction*, IEEE Transactions on Image Processing, 8 (1999), pp. 688–699.
[23] J. A. Fessler and B. P. Sutton, *Nonuniform fast fourier transforms using min-max interpolation*, IEEE transactions on signal processing, 51 (2003), pp. 560–574.
[24] A. K. Fletcher, S. Rangan, S. Sarkar, and P. Schniter, *Plug-in estimation in high-dimensional linear inverse problems: A rigorous analysis*, in NIPS, 2018.
[25] M. I. Florea and S. A. Vorobyov, *A generalized accelerated composite gradient method: Uniting nesterov's fast gradient method and fista*, arXiv preprint arXiv:1705.10266, (2017).
[26] B. Gonzales, D. Spronk, Y. Cheng, A. Tucker, M. Beckman, O. Zhou, and J. Lu, *Rectangular fixed-gantry CT prototype: combining cnt X-ray sources and accelerated compressed sensing-based reconstruction*, IEEE Access, 2 (2014), pp. 971–981.
[27] C. Guo and M. E. Davies, *Near optimal compressed sensing without priors: Parametric sure approximate message passing.*, IEEE Trans. Signal Processing, 63 (2015), pp. 2130–2141.
[28] J. Huang, Y. Zhang, J. Ma, D. Zeng, Z. Bian, S. Niu, Q. Feng, Z. Liang, and W. Chen, *Iterative image reconstruction for sparse-view CT using normal-dose image induced total variation prior*, PloS one, 8 (2013), p. e79709.
[29] J. H. Jørgensen, T. L. Jensen, P. C. Hansen, S. H. Jensen, E. Y. Sidky, and X. Pan, *Accelerated*

*gradient methods for total-variation-based CT image reconstruction*, arXiv:1105.4002, (2011).
[30] U. S. Kamilov, H. Mansour, and B. Wohlberg, *A plug-and-play priors approach for solving nonlinear imaging inverse problems*, IEEE Signal Processing Letters, 24 (2017), pp. 1872–1876.
[31] D. Kim, S. Ramani, and J. A. Fessler, *Combining ordered subsets and momentum for accelerated X-ray CT image reconstruction*, IEEE Transactions on Medical Imaging, 34 (2015), pp. 167–178.
[32] J. Ma and L. Ping, *Orthogonal amp*, IEEE Access, 5 (2017), pp. 2020–2033.
[33] A. Manoel, F. Krzakala, E. Tramel, and L. Zdeborova, *Swept approximate message passing for sparse estimation*, in International Conference on Machine Learning, 2015, pp. 1123–1132.
[34] S. Matej, J. A. Fessler, and I. Kazantsev, *Iterative tomographic image reconstruction using fourier-based forward and back-projectors*, IEEE transactions on medical imaging, 23 (2004), pp. 401–412.
[35] C. A. Metzler, A. Maleki, and R. G. Baraniuk, *From denoising to compressed sensing*, IEEE Transactions on Information Theory, 62 (2016), pp. 5117–5144.
[36] J. G. Nagy, R. J. Plemmons, and T. C. Torgersen, *Iterative image restoration using approximate inverse preconditioning*, IEEE Transactions on Image Processing, 5 (1996), pp. 1151–1162.
[37] M. Nilchian, C. Vonesch, P. Modregger, M. Stampanoni, and M. Unser, *Fast iterative reconstruction of differential phase contrast X-ray tomograms*, Optics express, 21 (2013), pp. 5511–5528.
[38] S. Niu, Y. Gao, Z. Bian, J. Huang, W. Chen, G. Yu, Z. Liang, and J. Ma, *Sparse-view X-ray CT reconstruction via total generalized variation regularization*, Physics in medicine and biology, 59 (2014), p. 2997.
[39] J. Nuyts, B. De Man, J. A. Fessler, W. Zbijewski, and F. J. Beekman, *Modelling the physics in the iterative reconstruction for transmission computed tomography*, Physics in medicine and biology, 58 (2013), p. R63.
[40] Y. O'Connor and J. A. Fessler, *Fourier-based forward and back-projectors in iterative fan-beam tomographic image reconstruction*, IEEE transactions on medical imaging, 25 (2006), pp. 582–589.
[41] S. Ono, *Primal-dual plug-and-play image restoration*, IEEE Signal Processing Letters, 24 (2017), pp. 1108–1112.
[42] M. Opper and O. Winther, *Expectation consistent approximate inference*, Journal of Machine Learning Research, 6 (2005), pp. 2177–2204.
[43] B. O'donoghue and E. Candes, *Adaptive restart for accelerated gradient schemes*, Foundations of computational mathematics, 15 (2015), pp. 715–732.
[44] A. Perelli and M. E. Davies, *Compressive computed tomography image reconstruction with denoising message passing algorithms*, in Signal Processing Conference (EUSIPCO), 2015 23rd European, IEEE, 2015, pp. 2806–2810.
[45] D. Potts and G. Steidl, *New fourier reconstruction algorithms for computerized tomography*, in Wavelet Applications in Signal and Image Processing VIII, vol. 4119, International Society for Optics and Photonics, 2000, pp. 13–24.
[46] S. Ramani, T. Blu, and M. Unser, *Monte-carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms*, IEEE Transactions on Image Processing, 17 (2008), pp. 1540–1554.
[47] S. Ramani and J. A. Fessler, *A splitting-based iterative algorithm for accelerated statistical X-ray CT reconstruction*, IEEE Transactions on Medical Imaging, 31 (2012), pp. 677–688.
[48] S. Ramani, X. Wang, L. Fu, and M. Lexa, *Denoising-based accelerated statistical iterative reconstruction for X-ray CT*, in 4th International Meeting on Image Formation in X-Ray Computed Tomography, 2016, pp. 337–340.
[49] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, arXiv preprint arXiv:1010.5141, (2010).
[50] S. Rangan, *Generalized approximate message passing for estimation with random linear mixing*, in IEEE International Symposium on Information Theory Proceedings, 2011, pp. 2168–2172.
[51] S. Rangan, A. K. Fletcher, P. Schniter, and U. S. Kamilov, *Inference for generalized linear models via alternating directions and Bethe free energy minimization*, IEEE Transactions on Information Theory, 63 (2016), pp. 676–697.
[52] S. Rangan, P. Schniter, and A. K. Fletcher, *Vector approximate message passing*, IEEE Transactions on Information Theory, 65 (2019), pp. 6664–6684.
[53] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar, *On the convergence of approximate mes-*

*sage passing with arbitrary matrices*, IEEE Transactions on Information Theory, 65 (2019), pp. 5339–5351.

[54] E. T. Reehorst and P. Schniter, *Regularization by denoising: Clarifications and new interpretations*, IEEE Transactions on Computational Imaging, 5 (2018), pp. 52–67.

[55] Y. Romano, M. Elad, and P. Milanfar, *The little engine that could: Regularization by denoising (red)*, SIAM Journal on Imaging Sciences, 10 (2017), pp. 1804–1844.

[56] S. Sarkar, A. K. Fletcher, S. Rangan, and P. Schniter, *Bilinear recovery using adaptive vector-amp*, IEEE Transactions on Signal Processing, 67 (2019), pp. 3383–3396.

[57] K. Sauer and C. Bouman, *A local update strategy for iterative reconstruction from projections*, IEEE Transactions on Signal Processing, 41 (1993), pp. 534–548.

[58] P. Schniter, S. Rangan, , and A. K. Fletcher, *Plug-and-play image recovery using vector amp*, in International BASP Frontiers workshop 2017, 2017, http://www2.ece.ohio-state.edu/~schniter/pdf/basp17_poster.pdf.

[59] P. Schniter and S. Rangan, *Compressive phase retrieval via generalized approximate message passing*, IEEE Transactions on Signal Processing, 63 (2015), pp. 1043–1055.

[60] P. Schniter, S. Rangan, and A. K. Fletcher, *Vector approximate message passing for the generalized linear model*, in 50th Asilomar Conference on Signals, Systems and Computers, 2016, pp. 1525–1529.

[61] M. Sharma, S. Metzler, C. A.and Nagesh, O. Cossairt, R. G. Baraniuk, and A. Veeraraghavan, *Inverse scattering via transmission matrices: Broadband illumination and fast phase retrieval algorithms*, IEEE Transactions on Computational Imaging, (2019).

[62] E. Y. Sidky, C.-M. Kao, and X. Pan, *Accurate image reconstruction from few-views and limited-angle data in divergent-beam ct*, Journal of X-ray Science and Technology, 14 (2006), pp. 119–139.

[63] S. Sreehari, S. V. Venkatakrishnan, and B. Wohlberg, *Plug-and-play priors for bright field electron tomography and sparse interpolation*, IEEE Transactions on Computational Imaging, 2 (2016), pp. 408–423.

[64] J. Tan, Y. Ma, H. Rueda, D. Baron, and G. R. Arce, *Compressive hyperspectral imaging via approximate message passing*, IEEE Journal of Selected Topics in Signal Processing, 10 (2016), pp. 389–401.

[65] L. Tierney and J. B. Kadane, *Accurate approximations for posterior moments and marginal densities*, Journal of the American statistical association, 81 (1986), pp. 82–86.

[66] W. van Aarle, W. J. Palenstijn, J. Cant, E. Janssens, F. Bleichrodt, A. Dabravolski, J. De Beenhouwer, K. J. Batenburg, and J. Sijbers, *Fast and flexible X-ray tomography using the astra toolbox*, Optics express, 24 (2016), pp. 25129–25147.

[67] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, *Plug-and-play priors for model based reconstruction*, in 2013 IEEE Global Conference on Signal and Information Processing, IEEE, 2013, pp. 945–948.

[68] J. Vila, P. Schniter, S. Rangan, F. Krzakala, and L. Zdeborová, *Adaptive damping and mean removal for the generalized approximate message passing algorithm*, in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 2021–2025.

[69] J. P. Vila and P. Schniter, *Expectation-maximization gaussian-mixture approximate message passing*, IEEE Transactions on Signal Processing, 61 (2013), pp. 4658–4672.

[70] J. P. Vila and P. Schniter, *An empirical-bayes approach to recovering linearly constrained non-negative sparse signals*, IEEE Transactions on Signal Processing, 62 (2014), pp. 4689–4703.

[71] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, *Image quality assessment: from error visibility to structural similarity*, IEEE transactions on image processing, 13 (2004), pp. 600–612.

[72] B. Whiting, P. Massoumzadeh, O. Earl, J. O'Sullivan, D. Snyder, and J. Williamson, *Properties of preprocessed sinogram data in X-ray computed tomography*, Medical physics, 33 (2006), pp. 3290–3303.

[73] Z. Yu, J.-B. Thibault, C. Bouman, K. Sauer, and J. Hsieh, *Fast model-based X-ray CT reconstruction using spatially nonhomogeneous ICD optimization*, IEEE Transactions on Image Processing, 20 (2011), pp. 161–175.

[74] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, *Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising*, IEEE Transactions on Image Processing, 26 (2017), pp. 3142–3155.