



# Distributed and parallel sparse convex optimization for radio interferometry with PURIFY

Luke Pratley<sup>a</sup>, Jason D. McEwen<sup>a</sup>, Mayeul d'Avezac<sup>b</sup>, Xiaohao Cai<sup>a</sup>, David Perez-Suarez<sup>b</sup>, Ilektra Christidi<sup>b</sup>, Roland Guichard<sup>b</sup>

<sup>a</sup>*Mullard Space Science Laboratory (MSSL), University College London (UCL), Holmbury St Mary, Surrey RH5 6NT, UK*

<sup>b</sup>*Research Software Development Group, Research IT Services, University College London (UCL), London WC1E 6BT, UK*

---

## Abstract

Next generation radio interferometric telescopes are entering an era of big data with extremely large data sets. While these telescopes can observe the sky in higher sensitivity and resolution than before, computational challenges in image reconstruction need to be overcome to realize the potential of forthcoming telescopes. New methods in sparse image reconstruction and convex optimization techniques (cf. compressive sensing) have shown to produce higher fidelity reconstructions of simulations and real observations than traditional methods. This article presents distributed and parallel algorithms and implementations to perform sparse image reconstruction, with significant practical considerations that are important for implementing these algorithms for Big Data. We benchmark the algorithms presented, showing that they are considerably faster than their serial equivalents. We then pre-sample gridding kernels to scale the distributed algorithms to larger data sizes, showing application times for 1 Gb to 2.4 Tb data sets over 25 to 100 nodes for up to 50 billion visibilities, and find that the run-times for the distributed algorithms range from 100 milliseconds to 3 minutes per iteration. This work presents an important step in working towards computationally scalable and efficient algorithms and implementations that are needed to image observations of both extended and compact sources from next generation radio interferometers such as the SKA. The algorithms are implemented in the latest versions of the SOPT (<https://github.com/astro-informatics/sopt>) and PURIFY (<https://github.com/astro-informatics/purify>) software packages (Versions 3.1.0), which have been released alongside of this article.

**Keywords:** techniques: interferometric, techniques: image processing, methods: data analysis

---

## 1. Introduction

Since the 1940's radio interferometric telescopes have enabled observations of the radio sky at higher resolution and sensitivity than ever possible using a single dish telescope [1]. By combining single dish radio telescopes into an array, the pairs of antenna directly sample Fourier coefficients of the radio sky [2]. The further apart the antenna, the higher the resolution, and the more antenna, the higher the sensitivity.

Next generation radio interferometric telescopes such as the Square Kilometre Array (SKA) will revolutionize astronomy. The high sensitivity and resolution of images from these telescopes are expected to enable new scientific discoveries in both cosmology and astrophysics [3]. Among the revolutionary science goals, these radio interferometric telescopes are designed to detect the star formation during the Epoch of Re-ionization (EoR) [4], and increase our understanding of the formation and evolution of the galaxy cluster environment and magnetic fields [5, 6].

However, radio interferometric telescopes have a limited number of antenna, meaning they cannot complete the Fourier domain of the radio sky, and therefore image reconstruction techniques need to be employed to estimate the true sky brightness distribution. This includes traditional methods such as CLEAN [7, 8, 9] and maximum entropy [10, 11], and state-of-the-art deconvolution methods such as sparse regularization algorithms [12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26] (cf. compressive sensing). Furthermore, measurements from all pairs of antenna will generate an enormous amount of data, with measurements varying in time and observational radio frequency; errors in the instrumental response also present a big data challenge [27]. For the SKA to achieve revolutionary science goals, it is critical to overcome revolutionary challenges in big data image reconstruction. In [15, 19], several algorithms from convex optimization were presented as tools for this big data challenge. Where it is suggested that these algorithms

can be distributed on a computing cluster for big data image reconstruction, while providing accurate representations of the radio sky, motivating the developments in this work.

To bridge the gap between research in new distributed convex optimization algorithms and application of these algorithms to real observations, the software packages SOPT and PURIFY have been developed in C++ for use by astronomers [15, 20]. SOPT hosts a collection of convex optimization algorithms that are used to solve inverse problems in signal processing. PURIFY hosts algorithms in radio interferometric imaging, and interfaces with SOPT to perform sparse image reconstruction<sup>1</sup> of radio interferometric observations.

PURIFY has been applied to simulated and real interferometric observations, and has been shown to produce sensitive, accurate and high-resolution models of the radio sky [15, 20, 21]. Additionally, unlike CLEAN reconstruction algorithms (a standard in radio astronomy) no restoration process is needed (e.g. combining of the reconstruction and residuals of the reconstruction as routinely performed with CLEAN), increasing the scientific merit of the reconstruction [20]. PURIFY was recently used in [21] to test and develop a new method of distributed wide-field interferometric image reconstruction, where wide-field and non-coplanar effects were modeled for over 17.5 million measurements from a real observation from the SKA precursor the Murchison Wide-Field Array (MWA) [28] using distributed convex optimization algorithms presented in this work. This method was recently extended to larger data sets from the MWA of over 100 million visibilities. We present the first implementation of a distributed convex optimization algorithm that uses MPI to reconstruct images from radio interferometric telescopes, which is now implemented in PURIFY. We build on the general distributed framework proposed by [19], but develop a number of alternative algorithms to better address practical considerations of MPI. This includes the distribution of Fourier transforms and images across multiple nodes, a global fidelity constraint, and the explicit description of the MPI communication processes at different stages of the algorithms. Furthermore, we consider the pre-sampling of gridding kernels to reduce memory overhead significantly. These developments are absolutely critical to realise an efficient distributed MPI implemented that can scale to big-data in practice. While we plan to make improvements to performance in the future, this article describes details of the distributed Alternating Direction Method of Multipliers (ADMM), degridding, and wavelet transform algorithms with the public release of SOPT and PURIFY software versions, which makes these new distributed methods available to astronomers.

This manuscript starts by reviewing the interferometric imaging measurement equation in Section 2. We then introduce sparse regularization for radio interferometric imaging in Section 3. In Section 4 we introduce some basics of proximal calculus. In Section 5 we introduce the serial Dual Forward-Backward based Alternating Direction Method of Multipliers (ADMM) algorithm. This sets the ground work for introducing computationally distributed wavelet and measurement operators and distributed ADMM algorithm in Section 6, which have recently been made publicly available in the software packages PURIFY (<https://github.com/astro-informatics/purify>) and SOPT (<https://github.com/astro-informatics/sopt>). These algorithms make use of degridding and gridding, wavelet transforms, and proximal operators to reconstruct high quality images of the radio sky while communicating data between compute nodes of a computing cluster using MPI. We demonstrate the implementations of the distributed algorithms in PURIFY in Sections 7, where we can distribute the data over a computing cluster. We use multi-threaded parallelization on a Graphics Processing Unit (GPU) or via OpenMP to parallelize across cores of a CPU node. Lastly, in Section 8 we show that pre-sampling the gridding kernel calculations can save memory and allow the application of ADMM to large data sets. We then show the application times for the distributed algorithms for up to 50 billion visibilities across the nodes of a computing cluster. We end with a conclusion in Section 9.

## 2. Radio Interferometric Measurement Equation

In this section we introduce concepts of radio interferometric imaging and the radio interferometric measurement equation. This measurement equation is calculated through the use of a linear measurement operator.

A radio interferometer is an array of radio antenna, where each pair of antenna results in a baseline and each baseline samples a Fourier coefficient of the brightness distribution across the radio sky. The separation of a baseline – the baseline length – determines the resolution scale of the Fourier coefficient. There are many measurement equations developed for radio interferometry [29, 30, 31, 32] that are based on different approximations and different levels of accuracy. The van Cittert-Zernike theorem [33] shows that for a small field of view and co-planar array, we find a simplified Fourier relation between visibilities and the sky brightness

$$y(u, v) = \int_{-1}^1 \int_{-1}^1 x(l, m) a(l, m) e^{-2\pi i(lu + mv)} dl dm, \quad (1)$$

where the measurements  $y$  are known as visibilities for the baseline separation  $(u, v)$ ,  $(l, m)$  are the sky coordinates,  $x$  is the sky brightness, and operator  $a$  includes direction dependent effects such as the primary beam and limits for

---

<sup>1</sup>Such algorithms have been used in the wider framework of compressive sensing.

the field of view. Equation (1) consists of linear operations, such as integration, multiplication, and convolution. Any linear operation has a matrix representation after choosing a fixed basis and coordinate system, which is particularly useful when working with the discretized version of equation (1).

To evaluate the measurement equation, and simulate the telescope, performing a Fourier transform is required. However, since  $(u, v)$  do not lie on a regular grid, it is not possible to use the speed of the standard Fast Fourier Transform (FFT). Instead, a non-uniform FFT is used to evaluate the measurement equation, where an FFT is applied, and the visibilities are interpolated off the FFT grid [34, 35]. This process is traditionally known as degridding (or sometimes a prediction step).

In a discrete setting, let  $\mathbf{x} \in \mathbb{R}^N$  and  $\mathbf{y} \in \mathbb{C}^M$  be the sky brightness and observed visibilities respectively. A non-uniform FFT can be represented by the following linear operations

$$\mathbf{y} = \mathbf{WGFZSx}, \quad (2)$$

where  $\mathbf{S}$  represents a gridding correction and modeling of measurement independent instrumental effects,  $\mathbf{Z}$  represents zero padding of the image (to increase the resolution FFT, upsampling the Fourier grid),  $\mathbf{F}$  is an FFT,  $\mathbf{G}$  represents a convolution matrix that interpolates measurements off the grid, and  $\mathbf{W}$  are noise weights applied to the measurements. These linear operators represent the application of the measurement equation. Let  $\Phi = \mathbf{WGFZS} \in \mathbb{C}^{M \times N}$ , which is typically called the measurement operator for  $N$  pixels and  $M$  measurements. This measurement operator is described in detail in [20]. Moreover, the linear measurement operator  $\Phi$  has its adjoint operator  $\Phi^\dagger$ , which, in practice, consists of applying these operators in reverse (i.e.  $\mathbf{S}^\dagger \mathbf{Z}^\dagger \mathbf{F}^\dagger \mathbf{G}^\dagger \mathbf{W}^\dagger$ ). In this work, we assume that  $\mathbf{y}$  are the weighted measurements, and the dirty map is defined as  $\Phi^\dagger \mathbf{y}$ .

### 3. Sparse Regularization

The previous section presents an inverse problem that has many solutions. Sparse regularization is a method that can estimate the radio sky brightness and isolate a single likely solution. In radio astronomy, the measurements have Gaussian uncertainty, leading to least squares minimization. To impose a penalty against over fitting of the radio sky, we can add a regularization term that penalizes models that over fit the measurements, i.e. a penalty that encourages the model to be sparse in parameters while fitting the radio sky. This leads to maximum a posteriori (MAP) estimation

$$\operatorname{argmin}_{\mathbf{x}} \gamma g(\mathbf{x}) + \|\mathbf{y} - \Phi \mathbf{x}\|_{\ell_2}, \quad (3)$$

where the least squares term is regularized by the function  $g$  and  $\gamma \geq 0$  determines the strength of the penalty against over fitting. The optimization problem presented with MAP estimation is known as the unconstrained problem. One challenge of using MAP estimation to perform sparse regularization is choosing a proper regularization parameter  $\gamma$  (although effective strategies do exist; [36]). The choice of  $\gamma$ , however, can be avoided after moving from the unconstrained problem in MAP estimation to the constrained problem

$$\operatorname{argmin}_{\mathbf{x}} g(\mathbf{x}), \quad \text{s.t.} \quad \|\mathbf{y} - \Phi \mathbf{x}\|_{\ell_2} \leq \epsilon, \quad (4)$$

where  $\epsilon$  is the error tolerance. One main advantage of the constrained objective function, compared to the unconstrained form (3), is that the parameter  $\epsilon$  can be estimated from  $\mathbf{y}$  [20], and therefore could be easier to set than assign a pertinent value for  $\gamma$  in (3). Note, in practice, that the weights in  $\mathbf{y}$  might be relative with no flux scale attached, or are not reliable, which will cause a difficulty for the constrained problem. On the other hand, progress is being made on methods that can estimate values of  $\gamma$  for the unconstrained problem. It is also worth noticing that these two forms, (3) and (4), have close relationship and, in some sense, are equivalent to each other after assigning proper values for  $\epsilon$  and  $\gamma$ . The remainder of this work is focused on the constrained problem (4) and we assume  $\epsilon$  can be estimated.

#### 3.1. Analysis and Synthesis

In the following we focus on using the  $\ell_1$ -norm for the function  $g$  and require our solution to have positive real values, where the  $\ell_p$ -norm is defined by  $\|\mathbf{x}\|_{\ell_p} = (\sum_i x_i^p)^{1/p}$  for  $p > 0$ . Additionally, we need to choose the representation of our signal to efficiently model the sky. This is done using a linear transform  $\Psi$ , with the convention that  $\mathbf{x} = \Psi \boldsymbol{\alpha}$ , where  $\boldsymbol{\alpha}$  represents the coefficients of  $\mathbf{x}$  under the basis or dictionary  $\Psi$ . A wavelet transform is convenient because it can efficiently represent structures as a function of scale and position. Moreover,  $\Psi$  is not restricted to be a basis, but can be an over-complete frame containing a collection of transforms. In this work, we use a collection of wavelet transforms to model the radio sky, as done in [37, 19, 20, 21].

The synthesis forms of the objective function for the unconstrained and constrained problems are respectively

$$\mathbf{x}^* = \Psi \times \operatorname{argmin}_{\alpha} \left\{ \|\mathbf{y} - \Phi \Psi \alpha\|_{\ell_2}^2 / 2\sigma^2 + \gamma \|\alpha\|_{\ell_1}, \quad \text{s.t. } \Psi \alpha \in \mathbb{R}_+ \right\}, \quad (5)$$

$$\mathbf{x}^* = \Psi \times \operatorname{argmin}_{\alpha} \left\{ \|\alpha\|_{\ell_1}, \quad \text{s.t. } \|\mathbf{y} - \Phi \Psi \alpha\|_{\ell_2} \leq \epsilon \quad \& \quad \Psi \alpha \in \mathbb{R}_+ \right\}. \quad (6)$$

The analysis forms of the objective function for the unconstrained and constrained problems are respectively

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \left\{ \|\mathbf{y} - \Phi \mathbf{x}\|_{\ell_2}^2 / 2\sigma^2 + \gamma \|\Psi^\dagger \mathbf{x}\|_{\ell_1}, \quad \text{s.t. } \mathbf{x} \in \mathbb{R}_+ \right\}, \quad (7)$$

$$\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} \left\{ \|\Psi^\dagger \mathbf{x}\|_{\ell_1}, \quad \text{s.t. } \|\mathbf{y} - \Phi \mathbf{x}\|_{\ell_2} \leq \epsilon \quad \& \quad \mathbf{x} \in \mathbb{R}_+ \right\}. \quad (8)$$

In the synthesis form we solve for the wavelet coefficients  $\alpha$  directly and in the analysis form we solve for the pixel coefficients  $\mathbf{x}$  directly. In practice they provide different results depending on the problem to be solved [38]. We follow the work of [37], which uses an over-complete frame in the analysis setting and is typically found to provide better reconstruction quality than the synthesis setting. The objective function can be solved multiple-times after reweighting the  $\ell_1$ -norm in the analysis setting with an over-complete frame, using what is called Sparsity Averaging Reweighted Analysis (SARA) [37].

Recent works have considered polarimetric [26, 39, 40] and spectral sparse image reconstruction [41, 42]. The works of [39, 40] show that where polarimetric images are reconstructed as a four component vector of Stokes parameters  $I$  (total intensity),  $Q$  and  $U$  (linear polarizations), and  $V$  (circular polarization), it is possible to enforce the physical constraint that  $I \geq \sqrt{Q^2 + U^2 + V^2}$ . Such a constraint enforces physical structures on both total intensity and polarized intensity, increasing the physicality of the reconstructions. Additionally, it is possible to impose non-parametric structures on spectra, such as spectral smoothness or sparsity, increasing the fidelity across the spectrum.

The challenge in finding the global solution of these objective functions, (5)–(8), is that they are non-differentiable (because of the non-differentiability of the  $\ell_1$  regularization term) and are not always continuous (because they contain constraints). However, these objective functions have the property that they are convex and lower semi-continuous (l.s.c.). In the following sections, we introduce proximal operators, which provide tools and algorithms that can be used to find solutions to the above convex minimization problems.

#### 4. Proximal Operators

In the previous section we introduced the convex objective functions (3) and (4), which need to be minimized to obtain a likely solution of the radio sky. When the problem is poised as minimization of a convex cost function, there are many convex optimization tools – proximal operators and proximal algorithms among them – on hand to solve it and find a global minimizer. In the following, we briefly recall some concepts and operators of convex functions and convex sets, which are useful when discussing solutions to convex inverse problems. A more detailed introduction to these concepts can be found in [43, 44, 45], and have been discussed in the context of radio interferometric imaging previously [15, 37, 18, 19, 20, 21]. In this section, we review the basic mathematics of proximal operators, and introduce the closed-form solution of proximal operators used in this work.

Let  $X$  be a vector space and  $\Gamma_0(X)$  be the class of proper, l.s.c. convex functions that map from  $X$  to  $(-\infty, +\infty]$ . A function  $h$  is convex when

$$h(\alpha \mathbf{x}_1 + (1 - \alpha) \mathbf{x}_2) \leq \alpha h(\mathbf{x}_1) + (1 - \alpha) h(\mathbf{x}_2), \quad \forall \mathbf{x}_1, \mathbf{x}_2 \in X, \forall \alpha \in [0, 1], \quad (9)$$

which is then true for  $\forall h \in \Gamma_0(X)$ . The conjugate of  $h \in \Gamma_0(X)$ , denoted by  $h^* \in \Gamma_0(X)$ , is defined as

$$h^*(\mathbf{m}) := \sup_{\mathbf{x} \in X} (\mathbf{m}^\top \mathbf{x} - h(\mathbf{x})). \quad (10)$$

The conjugate can be used to map a convex objective function from the primal representation to its dual representation, where both representations have the same optimal values when strong duality holds [43, 44, 46, 45]. The subdifferential of  $h$  at  $\mathbf{x} \in X$ , denoted by  $\partial h(\mathbf{x})$ , is defined as

$$\partial h(\mathbf{x}) := \{\mathbf{u} \in X : h(\mathbf{z}) \geq h(\mathbf{x}) + \mathbf{u}^\top (\mathbf{z} - \mathbf{x}), \forall \mathbf{z} \in X\}. \quad (11)$$

When  $h$  is differentiable, the subdifferential is a singleton containing the gradient  $\nabla h$ . If  $\mathbf{0} \in \partial h(\mathbf{x})$  then  $\mathbf{x}$  belongs to the set of global minimizers of  $h$  [44].

For  $\forall h \in \Gamma_0(X)$  and any constant  $\lambda > 0$ , the proximity operator of function  $\lambda h$  at  $\mathbf{v} \in X$ , which is denoted by  $\text{prox}_{\lambda h}(\mathbf{v})$  and maps between  $X \rightarrow X$ , is defined as the solution of the minimization problem

$$\text{prox}_{\lambda h}(\mathbf{v}) = \underset{\mathbf{x} \in X}{\operatorname{argmin}} \left( \lambda h(\mathbf{x}) + \frac{1}{2} \|\mathbf{x} - \mathbf{v}\|_{\ell_2}^2 \right). \quad (12)$$

We see that  $\text{prox}_{\lambda h}(\mathbf{v})$  is a point that is chosen in  $X$  by compromising between minimizing  $h$  and being close to  $\mathbf{v}$ , where this compromise is weighted by  $\lambda$ . For large  $\lambda$  more movement is taken towards minimizing  $h$ , and for small  $\lambda$  less movement is taken from  $\mathbf{v}$ . The proximal operator in (12) involves solving a minimization problem, which sometimes has a simple analytic form and sometimes not. When there is no analytic form it needs to be solved or estimated iteratively. It can be shown that the proximal operator is closely related to the subdifferential (11), being equivalent to the inverse operation  $(I + \lambda \partial h)^{-1}(\mathbf{v})$  [44].

When applied to a convex function, the proximal operator can be used to find a global minimizer through the recursive iteration. This is because the proximal operator is what is known as firmly non-expansive. More importantly it is a contraction, meaning repeated application of the proximal operator

$$\mathbf{x}^{k+1} = \text{prox}_{\lambda h}(\mathbf{x}^k) \quad (13)$$

will converge to a fixed point that minimizes  $\lambda h$  and therefore also minimizes  $h$ ; that is,  $\mathbf{x} = \text{prox}_{\lambda h}(\mathbf{x})$  if and only if  $\mathbf{x}$  minimizes  $h$  [43, 44].

The proximal operator has plenty of useful properties. For example, the proximal operator for the translation, the semi-orthogonal linear transform and the convex conjugation are

$$\text{prox}_{\lambda h(\cdot + \mathbf{a})}(\mathbf{x}) = \text{prox}_{\lambda h}(\mathbf{x} + \mathbf{a}) - \mathbf{a}, \quad \forall \mathbf{a} \in X, \quad (14)$$

$$\text{prox}_{\lambda h(\mathbf{L}(\cdot))}(\mathbf{x}) = \mathbf{x} + \mathbf{L}^\dagger (\text{prox}_{\lambda h}(\mathbf{Lx}) - \mathbf{Lx}), \quad \mathbf{L}\mathbf{L}^\dagger = \mathbf{I} \quad (15)$$

and

$$\text{prox}_{\lambda h^*}(\mathbf{x}) = \mathbf{x} - \lambda \text{prox}_{\lambda^{-1} h}(\mathbf{x}/\lambda), \quad (16)$$

respectively. The property for convex conjugation is also known as Moreau decomposition. Refer to [43, 44] and references therein for other properties and more details. Typically, it is difficult to obtain a closed form of the proximal operator for two functions  $f + g$ . The algorithms in the following section split the algorithm into solving for  $f + g$  given the proximal operator of  $f$  and  $g$  separately, and are typically called proximal splitting algorithms. First, we introduce closed forms of proximal operators that are used in radio interferometric imaging (but more examples are listed in [43, 44]).

In this work, we focus on  $\ell_1$  regularized least squares, i.e., using the  $\ell_1$  prior for  $g$  in the constrained problem (4). We need to minimize an  $\ell_1$ -norm with the condition that the solution lies within an  $\ell_2$ -ball with the size of our error  $\epsilon$ , while being real or positive valued. This can be mathematically stated as

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \|\Psi^\dagger \mathbf{x}\|_{\ell_1} + \iota_C(\mathbf{x}) + \iota_{B_{\ell_2}^\epsilon}(\Phi \mathbf{x}) \right\}, \quad (17)$$

where we normally take  $C = \mathbb{R}_+^N$ , and  $\ell_2$ -ball  $B_{\ell_2}^\epsilon$  to be the closed ball of radius  $\epsilon$ , and  $\iota_C(\mathbf{x})$  is the indicator function for  $\mathbf{x}$  being in  $C$  which will be detailed below. We now present the proximal operators needed to minimize this objective function.

#### 4.1. Indicator Function

Fix any nonempty closed convex set  $C$ , on which we define the indicator function as

$$\iota_C(\mathbf{x}) := \begin{cases} 0, & \mathbf{x} \in C, \\ +\infty, & \mathbf{x} \notin C. \end{cases} \quad (18)$$

We recall the projection operator  $\mathcal{P}_C$ , i.e.

$$\mathcal{P}_C(\mathbf{x}) := \underset{\mathbf{v} \in C}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{v}\|_{\ell_2}^2. \quad (19)$$

If  $C \subseteq X$ , then we have  $\iota_C \in \Gamma_0(X)$  and

$$\mathcal{P}_C(\mathbf{x}) = \underset{\mathbf{v} \in C}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{v}\|_{\ell_2}^2 = \underset{\mathbf{v} \in X}{\operatorname{argmin}} \left\{ \iota_C(\mathbf{v}) + \|\mathbf{x} - \mathbf{v}\|_{\ell_2}^2 \right\} = \text{prox}_{\iota_C}(\mathbf{x}). \quad (20)$$

Therefore, the proximal operator can be regarded as an extension of the projection operator [44]. The indicator function is useful for e.g. restricting a cost function to a set of solutions, or enforcing real or positive values on the solutions as assumptions for an image of the radio sky.

#### 4.2. Fidelity Constraint

Let the closed  $\ell_2$ -ball  $\mathcal{B}_{\ell_2}^\epsilon$  centered at  $\mathbf{z} \in X$  with radius  $\epsilon$  be the set

$$\mathcal{B}_{\ell_2}^\epsilon(\mathbf{z}) := \{\mathbf{v} \in X : \|\mathbf{z} - \mathbf{v}\|_{\ell_2} \leq \epsilon\}. \quad (21)$$

Then the proximal operator of an  $\ell_2$ -ball centered at zero reads

$$\begin{aligned} \text{prox}_{\mathcal{B}_{\ell_2}^\epsilon(0)}(\mathbf{x}) &= \underset{\mathbf{v} \in X}{\operatorname{argmin}} \left\{ \iota_{\mathcal{B}_{\ell_2}^\epsilon(0)}(\mathbf{v}) + \frac{1}{2} \|\mathbf{v} - \mathbf{x}\|_{\ell_2}^2 \right\} \\ &= \begin{cases} \mathbf{x}, & \mathbf{x} \in \mathcal{B}_{\ell_2}^\epsilon(0), \\ \frac{\mathbf{x}}{\|\mathbf{x}\|} \epsilon, & \mathbf{x} \notin \mathcal{B}_{\ell_2}^\epsilon(0). \end{cases} \end{aligned} \quad (22)$$

In detail, when  $\mathbf{x} \in \mathcal{B}_{\ell_2}^\epsilon(0)$ , we have  $\text{prox}_{\mathcal{B}_{\ell_2}^\epsilon(0)}(\mathbf{x}) = \mathbf{x}$  straightforwardly; when  $\mathbf{x} \notin \mathcal{B}_{\ell_2}^\epsilon(0)$ , computing  $\text{prox}_{\mathcal{B}_{\ell_2}^\epsilon(0)}(\mathbf{x})$  is to find a  $\mathbf{v} \in \mathcal{B}_{\ell_2}^\epsilon(\mathbf{0})$  such that it minimizes  $\|\mathbf{v} - \mathbf{x}\|_{\ell_2}^2$ . From the triangle inequality, we require that  $\mathbf{v}$  is parallel to  $\mathbf{x}$  for it to be a minimizer. It follows that we can scale  $\mathbf{x}$  into  $\mathcal{B}_{\ell_2}^\epsilon(0)$  to obtain the explicit representation of  $\text{prox}_{\mathcal{B}_{\ell_2}^\epsilon(0)}(\mathbf{x})$  shown in (22). Using the translation property of the proximal operator in (14), we can find the proximal operator of an  $\ell_2$ -ball centered at  $\mathbf{z}$ , i.e.,

$$\mathcal{P}_{\mathcal{B}}^\epsilon(\mathbf{z}) := \text{prox}_{\mathcal{B}_{\ell_2}^\epsilon(\mathbf{z})}(\mathbf{x}) = \begin{cases} \mathbf{x}, & \mathbf{x} - \mathbf{z} \in \mathcal{B}_{\ell_2}^\epsilon(\mathbf{0}), \\ \frac{\mathbf{x} - \mathbf{z}}{\|\mathbf{x} - \mathbf{z}\|} \epsilon + \mathbf{z}, & \mathbf{x} - \mathbf{z} \notin \mathcal{B}_{\ell_2}^\epsilon(\mathbf{0}). \end{cases} \quad (23)$$

#### 4.3. Promoting Sparsity

The  $\ell_1$ -norm is the sum of the absolute values of all components of a vector. Since it is convex and can promote sparsity when serving as a prior distribution or regularization, it is widely used in signal/image processing and has been shown highly effective in radio astronomy.

The proximal operator of the  $\ell_1$ -norm reads

$$\begin{aligned} \text{prox}_{\lambda \|\cdot\|_{\ell_1}}(\mathbf{x}) &= \underset{\mathbf{v} \in X}{\operatorname{argmin}} \left\{ \lambda \|\mathbf{v}\|_{\ell_1} + \frac{1}{2} \|\mathbf{v} - \mathbf{x}\|_{\ell_2}^2 \right\} \\ &= \mathcal{S}_\lambda(\mathbf{x}). \end{aligned} \quad (24)$$

Here  $\mathcal{S}_\lambda(\mathbf{x})$  is the soft thresholding of vector  $\mathbf{x} = (x_1, \dots, x_i, \dots)$ , which is defined as

$$\mathcal{S}_\lambda(\mathbf{x}) = (\mathcal{S}_\lambda(x_1), \dots, \mathcal{S}_\lambda(x_i), \dots), \quad (25)$$

where

$$\mathcal{S}_\lambda(x_i) = \begin{cases} 0, & |x_i| \leq \lambda, \\ \frac{x_i(|x_i| - \lambda)}{|x_i|}, & |x_i| > \lambda. \end{cases} \quad (26)$$

An intuitive derivation can be found by differentiating  $\lambda|v| + \frac{1}{2}(v - x)^2$  with respect to  $v$ , and then solving for

$$x = \begin{cases} 0, & v = 0, \\ v + \lambda & v > 0, \\ v - \lambda & v < 0. \end{cases} \quad (27)$$

Then case by case we find  $x = 0$  when  $v = 0$ ,  $x > \lambda$  when  $v > 0$  and  $x < -\lambda$  when  $v < 0$ . Rearranging the formula for each case, we arrive at soft thresholding as the proximal operator for the  $\ell_1$ -norm.

#### 4.4. Summary

This section has provided an introduction to proximal operators and examples of their closed-form solutions that are commonly used for interferometric imaging of real observations [19, 20, 22]. Proximal operators are especially powerful when the objective function is non-smooth, which is often required to enforce physicality on the solution. One important example in polarimetric imaging, but not detailed here, is where epigraphical projection techniques are used to construct a proximal operator that will project onto the set of solutions that contains  $I \geq \sqrt{U^2 + Q^2 + V^2}$  [39, 40].

We have provided proximal operators for a function  $f$ , but we often need to minimize an addition of functions, e.g.  $f + g$ . In the next section, we show how to solve for the minimizer of  $f + g$  when the proximal operators of  $f$  and  $g$  are known separately.

**Algorithm 1** Dual Forward-Backward ADMM.

The Dual Forward-Backward ADMM algorithm without MPI implementation. Lines 3–4 evaluate the  $\ell_2$ -ball proximal operator (constraining the solution to the  $\ell_2$ -ball), which is to address the solution of the subproblem (30). Line 5 is the Lagrangian dual variable update, connecting the two minimization problems (29) and (30). Lines 6–7 are a Forward-Backward step, which is to address the solution of the subproblem (29); particularly, line 6 is the forward (gradient) step, and line 7 is the backward step which is solved using the Dual Forward-Backward algorithm, as described between lines 9–16.

---

```

1: given  $\mathbf{x}^{(0)}, \mathbf{r}^{(0)}, \mathbf{s}^{(0)}, \mathbf{q}^{(0)}, \gamma, \rho, \varrho$ 
2: repeat for  $t = 1, \dots$ 
3:    $\mathbf{v}^{(t)} = \Phi\mathbf{x}^{(t-1)}$ 
4:    $\mathbf{r}^{(t)} = \mathcal{P}_{\mathcal{B}}^{\epsilon} \left( \mathbf{v}^{(t)} + \mathbf{s}^{(t-1)} \right)$ 
5:    $\mathbf{s}^{(t)} = \mathbf{s}^{(t-1)} + \varrho(\mathbf{v}^{(t)} - \mathbf{r}^{(t)})$ 
6:    $\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t-1)} - \rho\Phi^{\dagger} \left( \mathbf{v}^{(t)} - \mathbf{r}^{(t)} + \mathbf{s}^{(t)} \right)$ 
7:    $\mathbf{x}^{(t)} = \text{DUALFB}(\tilde{\mathbf{x}}^{(t)}, \gamma)$ 
8: until convergence
9: function DUALFB( $\mathbf{z}, \gamma$ )
10:  given  $\mathbf{d}_j^{(0)}, \eta$ 
11:   $\bar{\mathbf{z}}^{(0)} = \mathcal{P}_{\mathcal{C}}(\mathbf{z})$ 
12:  repeat for  $k = 1, \dots$ 
13:     $\mathbf{d}^{(k)} = \frac{1}{\eta} \left( \mathbf{I} - \mathcal{S}_{\gamma} \right) \left( \eta \mathbf{d}^{(k-1)} + \Psi^{\dagger} \bar{\mathbf{z}}^{(k-1)} \right)$ 
14:     $\bar{\mathbf{z}}^{(k)} = \mathcal{P}_{\mathcal{C}} \left( \mathbf{z}^{(k-1)} - \Psi \mathbf{d}^{(k)} \right)$ 
15:  until convergence
16: return  $\bar{\mathbf{z}}^{(k)}$ 
```

---

## 5. Sparse Regularization using Dual Forward-Backward ADMM

As mentioned in (17), the standard constrained radio interferometry solution with  $\ell_1$  (sparse) regularization can be stated as

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \|\Psi^{\dagger} \mathbf{x}\|_{\ell_1} + \iota_{\mathcal{C}}(\mathbf{x}) + \iota_{\mathcal{B}_{\ell_2}^{\epsilon}(\mathbf{y})}(\Phi \mathbf{x}) \right\}, \quad (28)$$

with  $\mathcal{B}_{\ell_2}^{\epsilon}(\mathbf{y}) = \{\mathbf{z} \in \mathbb{C}^M : \|\mathbf{z} - \mathbf{y}\|_{\ell_2} \leq \epsilon\}$  being the set that satisfies the fidelity constraint and  $\mathcal{C} = \mathbb{R}_+^N$  is the set that represents the positive and real constraint.

Let  $\mathbf{r}$  be the slack variable with the constraint  $\mathbf{r} = \Phi \mathbf{x}$ . As described in Section [Appendix A.3](#), to solve the above problem (28), ADMM can be applied by minimizing the Lagrangian of problem (28) corresponding to  $\mathbf{x}$  and  $\mathbf{r}$  alternatively, i.e.,

$$\min_{\mathbf{x}} \left\{ \mu \left[ \|\Psi^{\dagger} \mathbf{x}\|_{\ell_1} + \iota_{\mathcal{C}}(\mathbf{x}) \right] + \frac{1}{2} \|\Phi \mathbf{x} - (\mathbf{r} - \mathbf{s})\|_{\ell_2}^2 \right\}, \quad (29)$$

$$\min_{\mathbf{r}} \left\{ \mu \left[ \iota_{\mathcal{B}_{\ell_2}^{\epsilon}(\mathbf{y})}(\mathbf{r}) \right] + \frac{1}{2} \|\mathbf{r} - (\Phi \mathbf{x} + \mathbf{s})\|_{\ell_2}^2 \right\}, \quad (30)$$

where  $\mathbf{s}$  represents the Lagrangian multiplier. Algorithm 1 shows the Dual Forward-Backward ADMM algorithm used to solve problem (28). Recall that it is the same as the standard ADMM algorithm, but uses Dual Forward-Backward splitting with a Forward-Backward step to minimize the subproblem (29). The distributed version of this algorithm is presented in [19]. The serial version of this algorithm has been implemented in PURIFY and applied in [20] to simulated and real observations from radio interferometric telescopes previously.

## 6. Distributed Dual Forward-Backward ADMM

In the previous sections, we covered serial proximal optimization algorithms and serial operators. It is well known that these algorithms can be distributed (see [47, 45, 19] and references therein).

In the remainder of this work, we describe the details for how to modify these algorithms to be distributed over a computing cluster using the MPI standard commonly known as MPI. For clarity, we describe MPI implementations of operators in PURIFY and SOPT. The measurements and MPI processes are distributed across the nodes of a computing cluster.

### 6.1. MPI Framework

The MPI standard is a framework where multiple processes of the same program are run concurrently, communicating data and variables at sync points. This is commonly referred to as distributed memory parallelism. There are many independent processes (nodes) with their own data, but they can send messages containing data between them. This is different from the more typical shared memory parallelism, where a single process has access to all the data, but executes multiple threads for sections of the program (such as a loop with independent iterations). However, hybrids of shared and distributed memory parallelism are not uncommon, where nodes on a computing cluster send messages while performing multi-threaded operations. Please see [48] for a formal reference on MPI<sup>2</sup>.

The MPI framework contains a total number of processes  $n_d$ , each with a rank  $0 \leq j < n_d$ , all connected by a communicator for sending and receiving data. The most basic methods of a communicator consist of send and receive operations between individual processes. However, typically sending and receiving is performed in collective send and receive operations:

**Broadcast (one to many)** – Send a copy of a variable (scalar or array) from the root node to all nodes.

**Scatter (one to many)** – Scatter is where a root process contains an array; different sections of this array are sent to different nodes. The root process does not keep the sent data.

**Gather (many to one)** – Gather is where the root process receives data from all nodes. This could be sections of an array, or variables that are combined into an array on the root process.

**All to All (many to many)** – All to all is where data is communicated between all nodes at once. Each process sends and receives. This could be single variables or sections of arrays.

**Reduce (many to one)** – Reduce, or performing a reduction, is where a binary operation (assumed to be associative and commutative) is efficiently performed with a variable or array over the cluster with the result sent to the root process. Summation of variables across nodes is a common example of this. However, logical operations and max/min operations are also common.

**All reduce (many to many)** – All reduce is equivalent to a reduction, but the result is broadcasted to all nodes from the root process. All reduce with summation is called an all sum all operation.

The operation to broadcast a copy of  $\mathbf{x}$  onto each node can be represented by the linear operation

$$\begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_{n_d} \end{bmatrix} \mathbf{x} \quad (31)$$

where  $\mathbf{I}_j$  is an  $N \times N$  identity matrix. The adjoint of this operation is a reduction

$$\mathbf{x}_{\text{sum}} = [\mathbf{I}_1 \dots \mathbf{I}_{n_d}] \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_{n_d} \end{bmatrix}. \quad (32)$$

It is possible to view other MPI operations of sending data between nodes in the context of linear mappings.

### 6.2. Distributed Visibilities

The visibilities can be loaded on a root process then sorted into groups that are scattered to each node. This process splits and sorts the measurement vector  $\mathbf{y}$  into groups  $\mathbf{y}_j$ , where  $j$  is the rank of a process:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix}. \quad (33)$$

In this work, we sort the visibilities  $\mathbf{y}$  via ordering them by baseline length and dividing  $\mathbf{y}$  into sections of equal size  $\mathbf{y}_j$  to be scattered to each node. However, it is also possible to have each MPI process to read a different set of measurements. In principle, the weights and  $uvw$  coordinates are scattered with the visibilities.

If there is too much data to load the measurements onto one node, the data can be loaded in sections and then scattered to each node. After the data has been distributed, sorting into groups can be done using logical reductions, and then distributed to each node using an all to all operation. This has been done with the  $w$ -stacking algorithm in [21].

---

<sup>2</sup>Official versions of the MPI standard can be found online at <https://www.mpi-forum.org/docs/>.

### 6.3. Distributed Measurement Operator

For each group of visibilities  $\mathbf{y}_j$  on node  $j$ , there is a corresponding measurement operator  $\Phi_j$ . However, there are many ways to relate  $\Phi_j$  to the measurement operator for  $\mathbf{y}, \Phi$ ; we show two examples.

#### 6.3.1. Distributed Images

We can relate the MPI measurement operator to the serial operators by

$$\Phi = \begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_{n_d} \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_{n_d} \end{bmatrix}. \quad (34)$$

The forward operator can be expressed simply as independent measurement operators applied in parallel after broadcasting  $\mathbf{x}$ :

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix} = \begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_{n_d} \end{bmatrix} \begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_{n_d} \end{bmatrix} \mathbf{x}. \quad (35)$$

The adjoint operator can be expressed as the adjoint of independent measurement operators applied in parallel, followed by a reduction

$$\mathbf{x}_{\text{dirty}} = [\mathbf{I}_1 \dots \mathbf{I}_{n_d}] \begin{bmatrix} \Phi_1^\dagger & & \\ & \ddots & \\ & & \Phi_{n_d}^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix}. \quad (36)$$

However, with the MPI framework, it is efficient to always have a copy of the same image on each node so that other image domain operations can be performed in parallel (i.e. wavelet transforms). This can be ensured by combining the broadcast and reduction in a single all sum all operation during the adjoint. We work with the forward operator that applies each measurement operator independently on each node, with a copy of  $\mathbf{x}$  located on each node

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix} = \begin{bmatrix} \Phi_1 & & \\ & \ddots & \\ & & \Phi_{n_d} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \vdots \\ \mathbf{x} \end{bmatrix}, \quad (37)$$

and the adjoint operation can be performed by applying the adjoint of each measurement operator independently followed by an all sum all

$$\begin{bmatrix} \mathbf{x}_{\text{dirty}} \\ \vdots \\ \mathbf{x}_{\text{dirty}} \end{bmatrix} = [\mathbf{I}_1 \dots \mathbf{I}_{n_d}] \begin{bmatrix} \Phi_1^\dagger & & \\ & \ddots & \\ & & \Phi_{n_d}^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix}. \quad (38)$$

We can normalize the operator with the operator norm, by using the power method to estimate the largest eigenvalue, and remove arbitrary scaling due to  $n_d$  and other normalization factors. This implementation is unique to this work and is different from the method discussed next which is used in [19].

#### 6.3.2. Distributed FFT Grid Sections

Another method we implemented, which is discussed in [19], is to distribute the grid points of the FFT grid, where the degridding can be performed on each node. This can be performed using a scatter and gather operation from a root process. We can define the operation of distributing the necessary grid points using the operators  $\mathbf{M}_j \in \mathbb{R}^{B_j \times 2N}$ , where  $B_j$  is the number of non zero columns of  $\mathbf{G}_j$ . Additionally, we can remove the zero columns of  $\mathbf{G}_j$ , such that  $\mathbf{G}_j \in \mathbb{R}^{M_j \times B_j}$ .

The measurement operator is defined by

$$\begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_{n_d} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \mathbf{G}_1 & & \\ & \ddots & \\ & & \mathbf{W}_{n_d} \mathbf{G}_{n_d} \end{bmatrix} \begin{bmatrix} \mathbf{M}_1 \\ \vdots \\ \mathbf{M}_{n_d} \end{bmatrix} \mathbf{FZSx}. \quad (39)$$

$[\mathbf{M}_1^\top, \dots, \mathbf{M}_{n_d}^\top]^\top$  can be seen as scattering the FFT grid points from the root process to the other nodes. The adjoint can be seen as gathering and summing gridded FFT grid points to the root process. While this method appears to reduce communication, this has the disadvantage that the result of the adjoint ends up only on the root process. In practice, this means a broadcast is eventually required after the adjoint of this measurement operator so that further image domain operations can be performed in parallel.

#### 6.4. Distributed Wavelet Operator

The MPI wavelet operator can be distributed for each wavelet basis in the dictionary. Using the convention that  $\mathbf{x} = \Psi\boldsymbol{\alpha}$ , each wavelet representation can be arranged as

$$\boldsymbol{\alpha} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_{n_w} \end{bmatrix}, \quad (40)$$

for  $n_w$  wavelet transforms. From this definition, it follows that each inverse transform is performed independently with a reduction at the end

$$\Psi = [\mathbf{I}_1 \dots \mathbf{I}_{n_w}] \begin{bmatrix} \Psi_1 & & \\ & \ddots & \\ & & \Psi_{n_w} \end{bmatrix}. \quad (41)$$

However, like with the distributed image measurement operator, we combine the reduction and broadcasting as an all sum all. In practice, we use the forward operation

$$\begin{bmatrix} \mathbf{x} \\ \vdots \\ \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_1 \\ \vdots \\ \mathbf{I}_{n_w} \end{bmatrix} [\mathbf{I}_1 \dots \mathbf{I}_{n_w}] \begin{bmatrix} \Psi_1 & & \\ & \ddots & \\ & & \Psi_{n_w} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_{n_w} \end{bmatrix}. \quad (42)$$

The adjoint operation is

$$\begin{bmatrix} \boldsymbol{\alpha}_1 \\ \vdots \\ \boldsymbol{\alpha}_{n_d} \end{bmatrix} = \begin{bmatrix} \Psi_1^\dagger & & \\ & \ddots & \\ & & \Psi_{n_w}^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \vdots \\ \mathbf{x} \end{bmatrix}. \quad (43)$$

#### 6.5. Distributed Proximal Operator

The proximal operators for the  $\ell_1$ -norm,  $\ell_2$ -ball, and convergence criteria may require communication between nodes, which is discussed in this section.

##### 6.5.1. Sparsity and Positivity Constraint

The  $\ell_1$ -proximal norm does not need a communicator in itself. However,  $\Psi$  contains more than one wavelet transform. The proximal operator for the  $\ell_1$ -norm is solved iteratively using the Dual Forward-Backward method. The objective function that proximal operator minimizes can be computed to check that the iterations have converged. For a given  $\mathbf{x}$ , the proximal operator returns

$$\operatorname{argmin}_{\mathbf{z}} \left[ \iota_{\mathbb{R}_+^N}(\mathbf{z}) + \|\Psi^\dagger \mathbf{z}\|_{\ell_1} + \frac{1}{2\gamma} \|\mathbf{x} - \mathbf{z}\|_{\ell_2} \right]. \quad (44)$$

To assert that the Dual Forward-Backward method has converged to a minimum when calculating the proximal operator requires checking the variation of the  $\ell_1$ -norm; calculating the  $\ell_1$ -norm requires an MPI all sum all operation over wavelet coefficients. Another assertion that can be made is that the relative variation of  $\mathbf{x}$  is close to zero, which requires no communication.

##### 6.5.2. Fidelity Constraint

In the constrained minimization problem, the solution is constrained to be within the  $\ell_2$ -ball through the proximal operator  $\operatorname{prox}_{\mathcal{B}_{\ell_2}^\epsilon(\mathbf{y})}(\mathbf{v})$ . However, this proximal operator requires calculating the  $\ell_2$ -norm of the residuals  $\|\mathbf{v} - \mathbf{y}\|_{\ell_2}$ . When the visibilities are distributed on each node  $\mathbf{y}_i$ , this calculation requires an all sum all.

However, if each node constrains the solution to an independent local  $\ell_2$ -ball using  $\operatorname{prox}_{\mathcal{B}_{\ell_2}^{\epsilon_j}(\mathbf{y}_j)}(\mathbf{v}_j)$  with radius  $\epsilon_j$ , where  $\epsilon = \sqrt{\sum_{j=1}^{n_d} \epsilon_j^2}$ . This solution will also lie within the global  $\ell_2$ -ball where we have used  $\operatorname{prox}_{\mathcal{B}_{\ell_2}^\epsilon(\mathbf{y})}(\mathbf{v})$ , which can be shown using the triangle inequality. This requires less communication (introducing a new  $\epsilon_j$  for each node is suggested [19]). However, the communication overhead is negligible because much larger data rates are being communicated. Furthermore, using the global  $\ell_2$ -ball is more robust in convergence rate as it is independent of how the measurements are grouped across the nodes.

**Algorithm 2** Distributed Image (Dual Forward-Backward ADMM):

Every node has access to a global  $\ell_2$ -ball proximal and a serial version of the measurement operator  $\Phi_j$ . After the adjoint of the measurement operator is applied, an **AllSumAll** is performed over the returned image of each node  $j$ , then each node has the combined image. An **AllSumAll** is also used after the forward wavelet operator  $\Psi_j$ . Communication is needed in calculation of  $\mathcal{P}_B^\epsilon$  with an **AllSumAll** in the  $\ell_2$ -norm of the residuals. Using instead  $\mathcal{P}_{B_j}^{\epsilon_j}$  removes this communication overhead but changes the minimization problem.

---

```

1: given  $\mathbf{x}^{(0)}, \mathbf{r}_j^{(0)}, \mathbf{s}_j^{(0)}, \mathbf{q}_j^{(0)}, \gamma, \rho, \varrho$ 
2: repeat for  $t = 1, \dots$ 
3:    $\forall j \in \{1, \dots, n_d\}$  do in parallel
4:      $\mathbf{v}_j^{(t)} = \Phi_j \mathbf{x}^{(t-1)}$ 
5:      $\mathbf{r}_j^{(t)} = \mathcal{P}_B^\epsilon \left( \mathbf{v}_j^{(t)} + \mathbf{s}_j^{(t-1)} \right)$ 
6:      $\mathbf{s}_j^{(t)} = \mathbf{s}_j^{(t-1)} + \varrho (\mathbf{v}_j^{(t)} - \mathbf{r}_j^{(t)})$ 
7:      $\mathbf{q}_j^{(t)} = \Phi_j^\dagger \left( \mathbf{v}_j^{(t)} - \mathbf{r}_j^{(t)} + \mathbf{s}_j^{(t)} \right)$ 
8:      $\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t-1)} - \rho \text{AllSumAll}_j(\mathbf{q}_j^{(t)})$ 
9:      $\mathbf{x}^{(t)} = \text{DUALFB}(\tilde{\mathbf{x}}^{(t)}, \gamma)$ 
10:    end
11: until convergence
12: function DUALFB( $\mathbf{z}, \gamma$ )
13:   given  $\mathbf{d}_j^{(0)}, \eta$ 
14:    $\bar{\mathbf{z}}^{(0)} = \mathcal{P}_C(\mathbf{z})$ 
15:   repeat for  $k = 1, \dots$ 
16:      $\forall j \in \{1, \dots, n_w\}$  do in parallel
17:        $\mathbf{d}_j^{(k)} = \frac{1}{\eta} \left( \mathcal{I} - \mathcal{S}_\gamma \right) \left( \eta \mathbf{d}_j^{(k-1)} + \Psi_j^\dagger \bar{\mathbf{z}}^{(k-1)} \right)$ 
18:        $\bar{\mathbf{z}}^{(k)} = \mathcal{P}_C \left( \mathbf{z} - \text{AllSumAll}_j \left( \Psi_j \mathbf{d}_j^{(k)} \right) \right)$ 
19:     end
20:   until convergence
21: return  $\bar{\mathbf{z}}^{(k)}$ 
```

---

### 6.6. Distributed Convergence

There are multiple methods that can be used to check that the solution has converged. For example, when the relative difference of the solution between iterations is small, i.e.  $\|\mathbf{x}^{(i)} - \mathbf{x}^{(i-1)}\|_{\ell_2}/\|\mathbf{x}^{(i)}\|_{\ell_2} < \delta$  for a small  $\delta$ ; when the relative difference of the objective function between iterations is small; and the condition that the residuals of the solution lie within the  $\ell_2$ -ball<sup>3</sup>. These convergence criteria need to be communicated across the nodes. The convergence criteria need to be chosen carefully, since the quality of the output image can be degraded if the iterations have not converged sufficiently.

### 6.7. Distributed ADMM

With PURIFY, we build on the previous sections and combine the MPI distributed linear operators and proximal operators to solve the radio interferometric imaging inverse problem. The previous section discusses how to distribute  $\Phi$  for distributed  $\mathbf{y}_j$ , and how to distribute  $\Psi$  for distributed wavelet coefficients. In Algorithms 2 and 3, we outline MPI algorithms that use two variations of the measurement operator. Algorithm 2 uses an all sum all in the adjoint of the measurement operator following Section 6.3.1 and Algorithm 3 performs an FFT on the root node distributing parts of the grid following Section 6.3.2. In practice we recommend using Algorithm 2 as it can be easily modified to efficiently model wide-field effects, as demonstrated in [21]. Furthermore, Algorithm 2 is simpler to implement.

---

<sup>3</sup>A feature of ADMM is that it will not ensure that the residuals lie in the  $\ell_2$ -ball for each iteration but it will converge to this condition.

**Algorithm 3** Distributed Fourier Grid (Dual Forward-Backward ADMM):

Every node has access to a global  $\ell_2$ -ball proximal operator. The measurement operator is split. First, the root process computes  $\mathbf{FZS}$ , and scatters parts of the FFT grid  $\mathbf{b}_j^{(t)}$  to each node. Each node then applies  $\mathbf{G}_j$  to predict the visibilities for the  $j^{\text{th}}$  node. After the  $\ell_2$ -ball proximal operator is applied, each node applies  $\mathbf{G}_j^\dagger$ , then the root node gathers and adds the result. Then an update image is broadcast to the other nodes, which is needed for DUALFB( $\tilde{\mathbf{x}}^{(t)}, \gamma$ ). The rest of the algorithm is as in Algorithm 2.

---

```

1: given  $\mathbf{x}^{(0)}, \mathbf{r}_j^{(0)}, \mathbf{s}_j^{(0)}, \mathbf{q}_j^{(0)}, \gamma, \rho, \varrho$ 
2: repeat for  $t = 1, \dots$ 
3:    $\forall j \in \{1, \dots, n_{\text{d}}\}$  do in parallel
4:     Root process only:  $\tilde{\mathbf{b}}^{(t)} = \mathbf{FZS}\mathbf{x}^{(t-1)}$ 
5:      $\mathbf{b}_j^{(t)} = \mathbf{Scatter}_j(\mathbf{M}_j \tilde{\mathbf{b}}^{(t)})$ 
6:      $\mathbf{v}_j^{(t)} = \mathbf{G}_j \mathbf{b}_j^{(t)}$ 
7:      $\mathbf{r}_j^{(t)} = \mathcal{P}_{\mathcal{B}}^\epsilon \left( \mathbf{v}_j^{(t)} + \mathbf{s}_j^{(t-1)} \right)$ 
8:      $\mathbf{s}_j^{(t)} = \mathbf{s}_j^{(t-1)} + \varrho (\mathbf{v}_j^{(t)} - \mathbf{r}_j^{(t)})$ 
9:      $\mathbf{q}_j^{(t)} = \mathbf{G}_j^\dagger \left( \mathbf{v}_j^{(t)} - \mathbf{r}_j^{(t)} + \mathbf{s}_j^{(t)} \right)$ 
10:    Root process only:  $\mathbf{q}_j^{(t)} = \mathbf{Gather}_j(\mathbf{q}_j^{(t)})$ 
11:    Root process only:  $\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t-1)} - \rho \mathbf{Z}^\dagger \mathbf{F}^\dagger \sum_{j=1}^{n_{\text{d}}} \mathbf{M}_j^\dagger \mathbf{q}_j^{(t)}$ 
12:     $\tilde{\mathbf{x}}^{(t)} = \mathbf{Broadcast}(\tilde{\mathbf{x}}^{(t)})$ 
13:     $\mathbf{x}^{(t)} = \text{DUALFB}(\tilde{\mathbf{x}}^{(t)}, \gamma)$ 
14:  end
15: until convergence
16: function DUALFB( $\mathbf{z}, \gamma$ )
17:   given  $\mathbf{d}_i^{(0)}, \eta$ 
18:    $\bar{\mathbf{z}}^{(0)} = \mathcal{P}_{\mathcal{C}}(\mathbf{z})$ 
19:   repeat for  $k = 1, \dots$ 
20:      $\forall j \in \{1, \dots, n_{\text{w}}\}$  do in parallel
21:        $\mathbf{d}_j^{(k)} = \frac{1}{\eta} \left( \mathbf{I} - \mathbf{S}_\gamma \right) \left( \eta \mathbf{d}_j^{(k-1)} + \mathbf{\Psi}_j^\dagger \bar{\mathbf{z}}^{(k-1)} \right)$ 
22:        $\bar{\mathbf{z}}^{(k)} = \mathcal{P}_{\mathcal{C}} \left( \mathbf{z} - \text{AllSumAll}_j \left( \mathbf{\Psi}_j \mathbf{d}_j^{(k)} \right) \right)$ 
23:     end
24:   until convergence
25:   return  $\bar{\mathbf{z}}^{(k)}$ 

```

---

### 6.8. Global Fidelity Constraint ADMM

When the measurements are spread across the various nodes, communication is required to ensure that the same  $\ell_2$ -ball constraint is enforced across all measurements. The proximal operator for the  $\ell_2$ -ball is

$$\mathcal{P}_{\mathcal{B}}^\epsilon(\mathbf{z}_j) \triangleq \begin{cases} \epsilon \frac{\mathbf{z}_j - \mathbf{y}_j}{\sqrt{\text{AllSumAll}_j(\|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2}^2)}} + \mathbf{y}_j & \sqrt{\text{AllSumAll}_j(\|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2}^2)} > \epsilon \\ \mathbf{z}_j & \sqrt{\text{AllSumAll}_j(\|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2}^2)} \leq \epsilon \end{cases}, \quad (45)$$

The global  $\epsilon$  with a measurement operator is different from other work, and requires a communication step not described in other work.

### 6.9. Local Fidelity Constraint ADMM

We can split the single  $\ell_2$ -ball into many, and restate a new constrained problem, i.e.,

$$\mathbf{x}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ \|\mathbf{\Psi}^\dagger \mathbf{x}\|_{\ell_1} + \iota_{\mathcal{C}}(\mathbf{x}) + \sum_{j=1}^{n_{\text{d}}} \iota_{\mathcal{B}_j}(\mathbf{\Phi}_j \mathbf{x}) \right\}. \quad (46)$$

In particular, the alternating minimization involving the slack variable  $\mathbf{r}$  is split into solving each  $\mathbf{r}_j$  independently

$$\min_{\mathbf{r}_j} \left\{ \mu [\iota_{\mathcal{B}_j}(\mathbf{r}_j)] + \frac{1}{2} \|\mathbf{r}_j - \Phi_j \mathbf{x} - \mathbf{s}_j\|_{\ell_2}^2 \right\}. \quad (47)$$

Each  $\ell_2$ -ball proximal operator acts on a different section of  $\mathbf{y}_j$ , so they can be performed in parallel with no communication [19]

$$\mathcal{P}_{\mathcal{B}_j}^{\epsilon_j}(\mathbf{z}_j) \triangleq \begin{cases} \epsilon_j \frac{\mathbf{z}_j - \mathbf{y}_j}{\|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2}} + \mathbf{y}_j & \|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2} > \epsilon_j \\ \mathbf{z}_j & \|\mathbf{z}_j - \mathbf{y}_j\|_{\ell_2} \leq \epsilon_j \end{cases}. \quad (48)$$

By replacing  $\mathcal{P}_{\mathcal{B}}^{\epsilon}(\mathbf{z}_j)$  with  $\mathcal{P}_{\mathcal{B}_j}^{\epsilon_j}(\mathbf{z}_j)$  in Algorithms 2 and 3, the communication needed can be reduced.

The reduced communication overhead due to the local  $\ell_2$ -ball constraint is negligible compared to the overhead of the **Broadcast** and **AllSumAll** operations performed on  $\mathbf{x}$ , since  $n_d \ll N$ . Additionally, there is a drawback when the convergence is sensitive to the distribution of  $\mathbf{y}_j$ , which is not the case for the global  $\ell_2$ -ball ADMM. We thus advocate using the global fidelity constraint.

## 7. Algorithm Performance using PURIFY

We have implemented the MPI ADMM algorithms from the previous sections in PURIFY and SOPT. In this section, we benchmark the performance against the non-distributed counterpart [20], to show that such methods can decrease the time required for each iteration. We also implement and benchmark a GPU implementation of the measurement operator against its CPU implementation, to show that GPU implementations can further increase the performance (which can be used in conjunction with the MPI algorithms).

### 7.1. PURIFY Software Package

PURIFY has been developed as a software package that will perform distributed sparse image reconstruction of radio interferometric observations to reconstruct a model of the radio sky. The sparse convex optimization algorithms and MPI operations have been implemented in a standalone library known as SOPT. Previous versions of PURIFY are described in [15, 20]. In this section, we describe the latest release of PURIFY (Version 3.1.0) [49] and latest release of SOPT (Version 3.1.0) [50] that accompany this article. You can download and find the latest details on PURIFY at <https://github.com/astro-informatics/purify> and SOPT at <https://github.com/astro-informatics/sopt>.

PURIFY and SOPT have been developed using the C++11 standard. We use the software package Eigen for linear algebra operations [51]. OpenMP is used to increase performance of the FFT, discrete planar wavelet transforms, and sparse matrix multiplication. The separable 2d discrete Daubechies wavelet transforms (1 to 30) have been implemented using a lifting scheme (more details on wavelet transforms can be found in [52, 53]), and have been multi-threaded over application of filters. The sparse matrix multiplication is multi-threaded over rows, requiring the sparse matrix to be stored in row-major order for best performance. To perform operations on a GPU, we use the library ArrayFire, which can be used with a CPU, OpenCL, or CUDA back-end [54]. Within SOPT, we have implemented various MPI functionality (all sum all, broadcast, gather, scatter, etc.) to interface with data types and communicate the algorithm operations across the cluster. It is possible to read the measurements (and associated data) using UVFITS or CASA Measurement Set (MS) formats. The UVFITS format follows the format set by [55]. The output images are saved as two dimensional FITS file images, as a sine projection. Currently, the distributed algorithm supported is ADMM. Furthermore,  $w$ -projection and  $w$ -stacking algorithms are supported for wide-fields of view and are described in [21].

### 7.2. Distribution of Visibilities

PURIFY can read visibilities  $\{\mathbf{y}_i\}_{i=1}^{n_d}$ , and scatter them to each node on the cluster. How these measurements are distributed is not important when using the global  $\ell_2$ -ball constraint. However, when using local  $\ell_2$ -ball constraints, the way the visibilities are grouped for each node could make a difference to convergence, where it could be better to keep similar baselines on the same node. We do this by assigning different nodes to different regions of the FFT grid, or by assigning different nodes to regions in baseline length  $\sqrt{u^2 + v^2}$ . However, when using the  $w$ -stacking algorithm  $k$ -means with MPI is used to redistribute the visibilities over the cluster using an all to all operation, as discussed in [21].

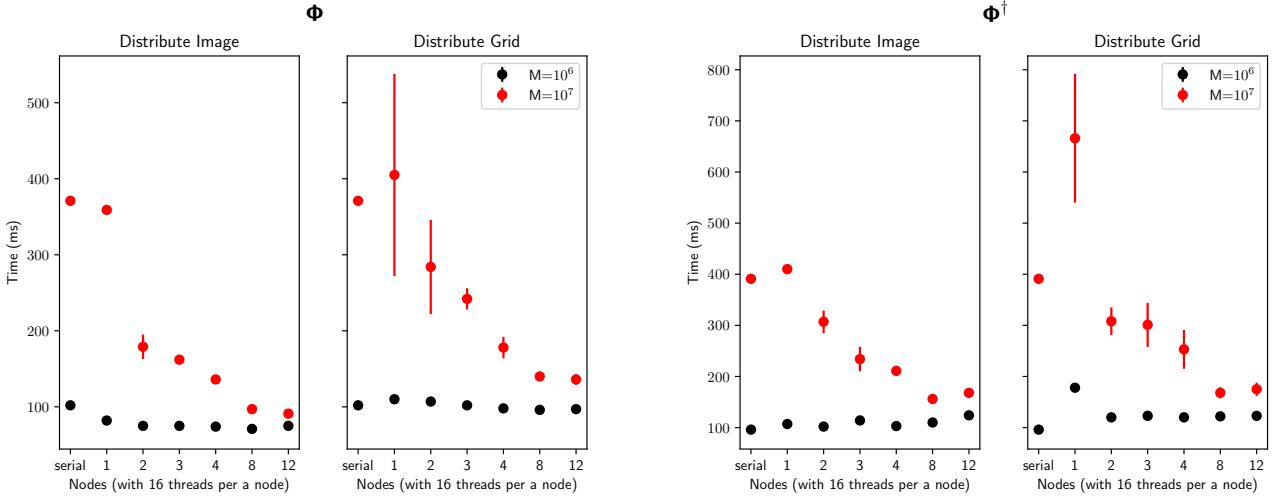


Figure 1. Time to apply forward  $\Phi$  (left two plots) and adjoint  $\Phi^\dagger$  (right two plots) as a function of the number of MPI nodes, benchmarked against the non MPI (serial) implementation. We fix the number of visibilities and image size at  $N = 1024 \times 1024$ ,  $M \in \{10^6, 10^7\}$ . On the left the MPI implementation corresponds to using an all sum all MPI operation in the adjoint described in Section 6.3.1; on the left MPI implementation corresponds to distribution of the grid from the root node, as described in Section 6.3.2.

### 7.3. Benchmark Timings

In the remainder of this section, we time the operations of the MPI algorithms. We use Google Benchmark<sup>4</sup> to perform the timings of the mean and standard deviation for each operation benchmarked. The times provided are in real time (incorporating communication), not CPU time, since multi-threaded operations are sensitive to this difference. Each benchmark configuration was timed for 10 runs, providing a mean and standard deviation used for timings and errors in the sections that follow.

The computing cluster Legion at University College London was used to measure the benchmark timings. We used the Type U nodes on Legion, which work as a 16 core device with 64GB RAM (160 Dell C6220 nodes - dual processor, eight cores per processor<sup>5</sup>).

In the benchmarks, the root node generates a random Gaussian density sampling distribution of baselines  $(u, v)$ , ranging from  $\pm\pi$  along each axis. The weights  $\mathbf{W}_j$  and baseline coordinates  $(\mathbf{u}_j, \mathbf{v}_j)$  are distributed to nodes  $1 \leq j \leq n_d$ . This allows the construction of  $\mathbf{W}_j \mathbf{G}_j$  on each node. We use the Kaiser-Bessel kernel as the interpolation (anti-aliasing) convolution kernel for  $\mathbf{G}_j$ , with a minimum support size of  $J = 4$  (see [20] for more details). The identical construction of  $\mathbf{FZS}$  can then be performed on each node or the root node (depending on the algorithm), and allow us to apply  $\Phi$  in each of the MPI algorithms.

### 7.4. MPI Measurement Operator Benchmarks

The **AllSumAll**( $\mathbf{x}$ ) and **Broadcast**( $\mathbf{x}$ ) in the  $\Phi^\dagger$  operations will be expensive in communication overheads for large  $N$ . Additionally, the calculation of the FFT  $\mathbf{F}$  does not take advantage of MPI and will have the cost  $\mathcal{O}(N \log N)$ , albeit the FFT is multi threaded using FFTW and OpenMP to provide performance improvements. It is more likely that the time taken to compute the FFT will take longer than the communication of the image at large  $N$ . If we evenly distribute the visibilities so that each node has  $M_j = M/n_d$ , the computational complexity of the sparse matrix multiplication  $\mathbf{G}_j$  reduces to  $\mathcal{O}(M_j J^2)$  per node, providing a large advantage at large  $M$  and  $n_d$ .

We benchmark the MPI  $\Phi$  and  $\Phi^\dagger$  implementations against the non-distributed equivalent using PURIFY. We use  $10^6$  and  $10^7$  visibilities, and a fixed image size of  $N = 1024 \times 1024$ . We vary the number of nodes from 1, 2, 3, 4, 8, 12. Results are shown in Figure 1. For  $10^6$  visibilities there is no improvement on the measurement operator performance for each MPI implementation. However, for  $10^7$  it is clear that increasing the number of nodes increases the performance. The saturation for  $n_d \geq 5$  can be explained by the computational cost of the FFT  $\mathbf{F}$  being greater than the sparse matrix multiplication  $\mathbf{G}_j$ . For small numbers of nodes  $n_d$ , i.e. 1 or 2, the application time is less reliable, but for larger  $n_d$  it becomes more stable. We also find that distributing sections of the grid (described in Section 6.3.2) is more expensive at low  $n_d$  than distributing the image (described in Section 6.3.1).

<sup>4</sup><https://github.com/google/benchmark>

<sup>5</sup>More details can be found at [https://wiki.rc.ucl.ac.uk/wiki/RC\\_Systems#Legion\\_technical\\_specs](https://wiki.rc.ucl.ac.uk/wiki/RC_Systems#Legion_technical_specs).

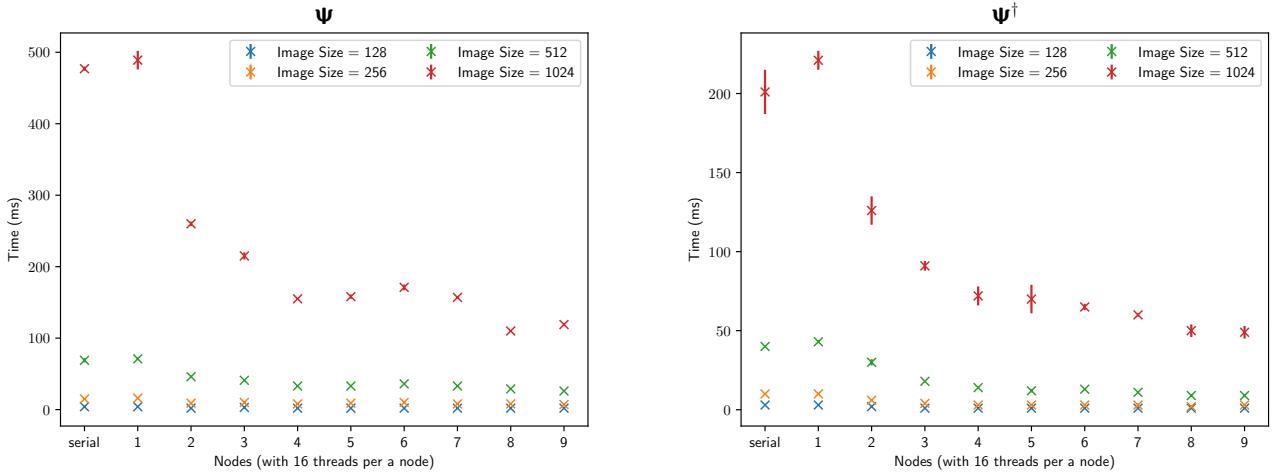


Figure 2. Time to apply forward  $\Psi$  (left) and adjoint  $\Psi^\dagger$  (right) as a function of the number of MPI nodes, benchmarked against the non MPI (serial) implementation. The forward operator requires 2 times more calculations than the adjoint due to the up sampling operations. Distributing the wavelet transforms across the nodes greatly decreases the time for calculation.

### 7.5. MPI Wavelet Operator Benchmarks

Like the measurement operator  $\Phi$ , the wavelet operator  $\Psi$  requires an **AllSumAll( $x$ )** operation. However, even with multi-threaded operations in the wavelet transform, computing  $\Psi_j$  is time consuming. When  $n_w > n_d$ , multiple wavelet transforms are performed on some of the nodes, for example when  $n_w = 2n_d$  there are two wavelet transforms per node. In many cases we expect that the numbers of nodes is greater than the number of wavelet transforms, i.e.  $n_d \geq n_w$ , and the maximum benefit from MPI distribution of wavelet transforms can be seen. This can be seen in Figure 2, where there is a performance improvement with using MPI to distribute the wavelet transforms across nodes.

In the benchmarks, we use  $n_w = 9$  where  $\Psi_0$  is a Dirac basis and  $\Psi_1$  to  $\Psi_8$  are 2d (the product of 1d) Db wavelets 1 to 8. We perform the wavelet transform to three levels. Increasing the number of wavelet levels requires more computation, but much of this computation is in the first few levels. Furthermore, the low pass and high pass filters in the Db increase with size from 1 to 8, meaning Db 8 requires more computation than Db 7 at each wavelet level (but we have found the time difference small). The forward operator  $\Psi$  requires up-sampling, meaning that it requires a factor of 2 times more computation than the adjoint  $\Psi^\dagger$ .

### 7.6. MPI Algorithm Benchmarks

As a demonstration the impact of the MPI operators, we benchmark the Algorithms 2 and Algorithm 3 against the serial Algorithm 1 (equivalent to  $n_d = 1$ ). We fix the number of visibilities and image size at  $N = 1024 \times 1024$ ,  $M \in \{10^6, 10^7\}$ .

We use local  $\ell_2$ -ball constraints for each node as described in Section 6.9. However, PURIFY also provides the ability to use the global  $\ell_2$ -constraint. In practice, we do not find much difference in computation time between using a local or global  $\ell_2$ -constraint.

In Figure 3, we time the application of one iteration of ADMM using one Dual Forward-Backward iteration. We find a clear increase in performance when increasing the number of nodes used. This is predicted from the performance improvements from the previous sections.

### 7.7. GPU Measurement Operator Benchmarks

The MPI measurement operators in the previous subsection can also make use of graphics processing units (GPUs) to increase performance. We have implemented the MPI measurement operators using the software package ArrayFire [54], which provides the flexibility to chose a CPU, CUDA, or OpenCL back-end to perform computations. The hybrid MPI-GPU measurement operator works the same as the MPI measurement operator, but all operations on a given node are performed on a GPU. In this section, we show that the GPU can increase performance. We benchmark the ArrayFire implementation using a CUDA back-end, against the equivalent measurement operator. No MPI is used in these benchmarks, since it is clear from the previous section that MPI will also increase performance. We perform the benchmarks on a high performance workstation, using an NVIDIA Quadro K4200 GPU (with 4GB RAM). We use  $5 \times 10^6$  visibilities, and use the image sizes of  $256 \times 256$ ,  $512 \times 512$ ,  $1024 \times 1024$  and  $2048 \times 2048$ . We find that the GPU implementation of degridding and gridding is about 10 times faster than the CPU counter part. Figure 4 shows the application of  $\Phi$  and  $\Phi^\dagger$ , with a large performance improvement when using the GPU.

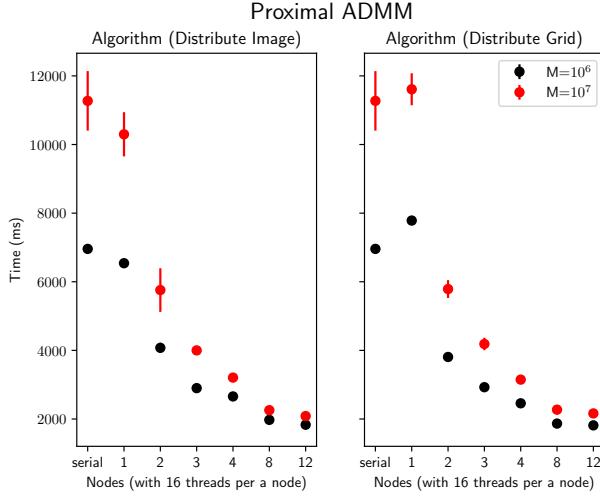


Figure 3. Time to apply a single iteration of the ADMM algorithm as a function of the number of MPI nodes, benchmarked against the non MPI (serial) implementation. The Dual Forward Backward algorithm is limited to one iteration. We fix the number of visibilities and image size at  $N = 1024 \times 1024$ ,  $M \in \{10^6, 10^7\}$ . On the left the MPI implementation corresponds to using Algorithm 2 (which uses the MPI measurement operator from Section 6.3.1 where the image is distributed); on the right MPI implementation corresponds to using Algorithm 3 (which uses the MPI measurement operator from Section 6.3.2 where the Fourier grid is distributed).

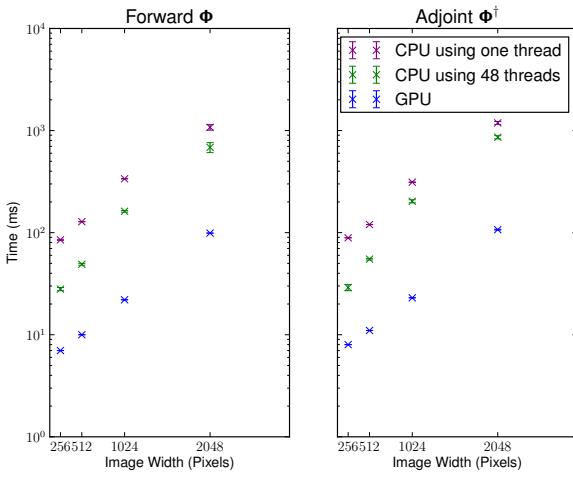


Figure 4. Time to apply forward  $\Phi$  and adjoint  $\Phi^\dagger$  as a function of image size, using CPU implementation and ArrayFire with GPU CUDA back-end implementation. We fix the number of visibilities at  $M = 5 \times 10^6$ , and vary the width of a square image. The CPU times for 1 and 48 threads show that there is some improvement by using threading for the CPU. However, it is clear that GPU implementation remains almost an order of magnitude faster for both gridding and degridding, especially at larger image sizes.

## 8. Application to Big-Data

In the previous section, the benchmarks against the serial version of the algorithm show that the distributed version of the algorithms can in principle increase the run-time speed. In this section, we pre-sample the gridding kernel onto a fine one dimensional grid to compress the storage of the gridding matrix. Since many of the kernel coefficients are reused, this can greatly reduce the memory needed during operation. We then use the pre-sampled kernel to time the distributed measurement operator and ADMM algorithms for larger data sets.

### 8.1. Kernel Pre-Sampling

The amount of memory required to store the sparse matrix  $\mathbf{G}$  can be quite substantial. When stored in the most naive way, it will take up  $J^2 M$  times the number of coefficients. This number increases by a factor of 2 if  $\mathbf{G}^\dagger$  is stored separately to ensure row major order during application. In this section, we discuss our strategy for compressing  $\mathbf{G}$  by using pre-sampling of the gridding kernel  $g(u_{\text{pix}})$  onto a fine grid over its support. Zeroth order interpolation can then be applied to calculate  $\mathbf{G}$  and  $\mathbf{G}^\dagger$  during application. When  $J = 4$  we find that we need to store 16 times as many weights as we do visibilities, which can cause us reach a memory bottleneck due to the limits of memory of the total cluster. It is easy to see that for  $J = 4$  and  $M = 100 \times 10^6$  will take up at least 25.6 Gb in complex valued weights, and this will double if  $\mathbf{G}^\dagger$  is stored separately.

On the fly calculation evaluation of  $g(u_{\text{pix}})$  frees up memory, but will take up calculation time during the application of  $\mathbf{G}$ . We have found that on the fly calculation time can be reduced by pre-sampling the kernel calculations.

This is not the first time that the idea of kernel pre-sampling has been used. The work of [56] discusses how pre-sampling can be used to accurately evaluate gridding kernels. In particular, they investigate the use of zero and first order interpolation.

In our work, we pre-sample  $g(u_{\text{pix}})$  for values between  $0 \leq |u_{\text{pix}}| \leq J/2$ . We use the knowledge that  $g(u_{\text{pix}}) = g(-u_{\text{pix}})$ , which halves the number of pre-samples needed. We have found that pre-sampling  $g(u_{\text{pix}})$  to a density of  $2 \times 10^5$  samples per pixel is sufficient to make zeroth order interpolation accurate to  $10^{-5}$ . This means that for  $J = 4$ , we need  $4 \times 10^5$  samples with the advantage that the number of pre-samples is independent of the number of measurements  $M$ . Furthermore, we can use the same pre-samples for the adjoint operator  $\mathbf{G}^\dagger$ .

In the next sections, we show the application of the distributed measurement operator and ADMM algorithms to larger data sets. In particular, pre-sampling the kernel calculation unlocks more computational resources during reconstruction of large data sets.

### 8.2. Measurement Operator and ADMM Scaling

Section 7 shows benchmarks against the serial algorithm, which shows that the distributed algorithms reduce the run-time (by increasing the available computational resources). However, benchmarking against a serial version is not possible or useful for understanding how the performance scales for very large data sets. We pre-sample the gridding kernels in order to time the application of the distributed algorithms to larger data sets.

To measure the timings, we fix the image size to  $N = 1024 \times 1024$  pixels and vary the number of visibilities per node  $M_j$ . The timings use the Grace computing cluster at University College London. Each node of Grace contains two 8 core Intel Xeon E5-2630v3 processors (16 cores total) and 64 Gb of RAM.<sup>6</sup> It is important to note that the  $uv$  sampling pattern has no clear impact on the time per iteration, especially with large data sizes.

Figures 5 (measurement operator) and 6 (ADMM) show how the application run-time varies as a function of visibilities per node  $M_j$ , using  $n_d = 25, 50, 100$  nodes. We did not find a large impact on performance when increasing the number of nodes for the different types of measurement operators and ADMM algorithms. This is important, because it suggests that communication overhead and lag in computation across nodes is currently not a major limitation or factor for the implementation of this algorithm up to 100 nodes.

Figure 6 also shows the total size of the data sets in memory across the cluster as a function of the number of nodes and visibilities per node. We define the total data set as number of complex valued visibilities  $\mathbf{y}_k$ , the complex valued weights  $\mathbf{W}_{kk}$ , and real valued  $\mathbf{u}_k$  and  $\mathbf{v}_k$  coordinates. We are using the data types ‘double’ (8 bytes) and ‘complex double’ (16 bytes), this can be multiplied by  $M$  or  $M_j$  to calculate the total memory over the cluster or per node respectively. The run-times are measured for data sets ranging from 1 Gb to 2.4 Tb. The largest data sets contain 12.5 billion, 25 billion and 50 billion visibilities, with  $M_j = 500 \times 10^6$  visibilities per node. The run-times for ADMM range from 9 seconds to 3 minutes per iteration, with a change in behavior for  $M_j > 10^7$ . However, the run-times for the measurement operators range from 100 milliseconds to 40 seconds per iteration.

---

<sup>6</sup>More details can be found at [https://wiki.rc.ucl.ac.uk/wiki/RC\\_Systems#Grace\\_technical\\_specs](https://wiki.rc.ucl.ac.uk/wiki/RC_Systems#Grace_technical_specs)

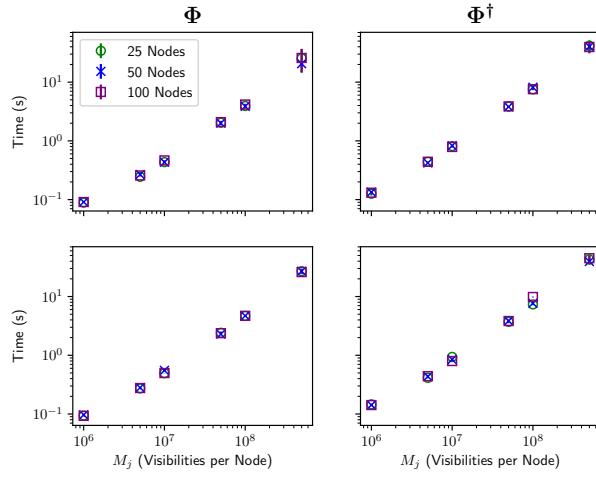


Figure 5. Application times of the distributed image (top row) and distributed grid measurement operators (left column) and their adjoint (right column) as a function of visibilities per node  $M_j$ . We show the application times for  $n_d = 25, 50, 100$  nodes. We find that application time increases linearly as a function of  $M_j$ . Importantly, choosing  $n_d = 100$  does not have a big impact on application time. The run-times in this figure measured for data sets the ranging from 1 Gb to 2.4 Tb.

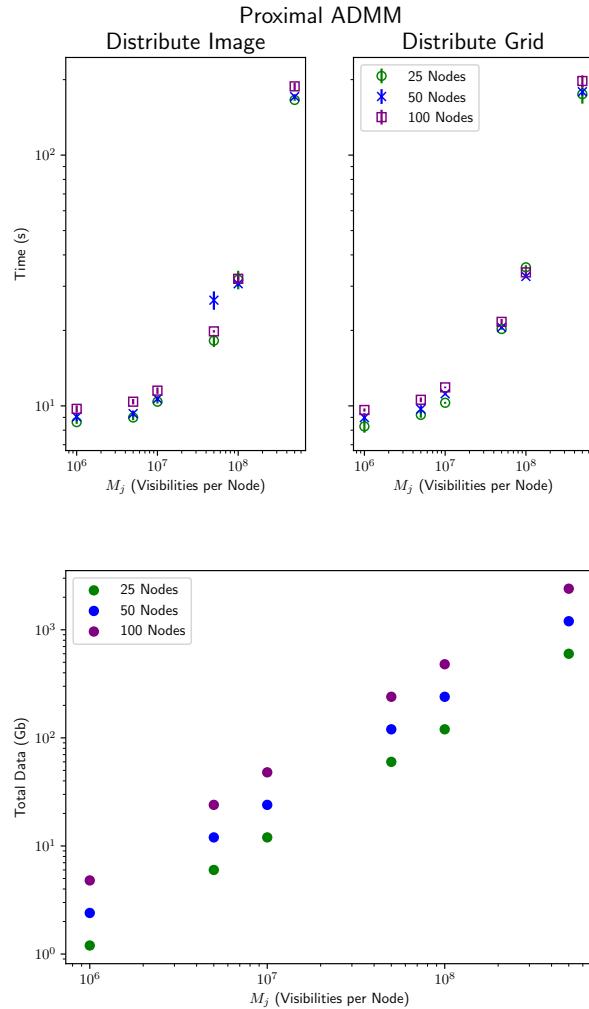


Figure 6. (top) Application times of the distributed image (left) and distributed grid ADMM algorithms (right) as a function of visibilities per node  $M_j$ . We show the application times for  $n_d = 25, 50, 100$  nodes. When  $M_j > 10^7$ , we find that scaling of ADMM changes in how it increases with  $M_j$ . (Bottom) The total size of data sets used in Figures 5 and 6. The data sets range from 1 Gb to 2.4 Tb.

## 9. Conclusion

We presented the first implementation of a distributed convex optimization algorithm that uses MPI to reconstruct images from radio interferometric telescopes. We have described new algorithms and their application to convex optimization methods for distributed interferometric image reconstruction as implemented by the PURIFY 3.1.0 and SOPT 3.1.0 software packages [49, 50]. We developed a number of alternative algorithms to better address practical considerations for MPI and include explicit descriptions of the MPI processes used. We then benchmark the distributed implementations to demonstrate considerable computational gains compared to the serial equivalents. We found that the pre-sampled gridding kernel calculation allows the distributed gridding operations and the ADMM algorithm to scale to big data sets that are over 1 Tb while providing a computational speed per iteration that does not have a strong communication overhead for up to 100 nodes. While we plan to make improvements to performance in the future, this work is a significant step towards new interferometric image reconstruction algorithms that can be applied to large data sets.

With next generation radio interferometric telescopes coming online, distributed and parallel image reconstruction and data analysis will be necessary to deal with the large image sizes and large volumes of data of forthcoming telescopes. This work is an important step on the path to developing computational algorithms that will be required for telescopes to reach the high resolution and sensitivity needed for science goals of telescopes such as the SKA.

## Acknowledgements

The authors thank Rafael Carrillo, Alex Onose and Yves Wiaux for useful discussions regarding the algorithms presented in [19]. This work was also supported by the Engineering and Physical Sciences Research Council (EPSRC) through grant EP/M011089/1 and by the Leverhulme Trust. The authors acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and associated support services, in the completion of this work. The authors acknowledge the use of the UCL Grace High Performance Computing Facility (Grace@UCL), and associated support services, in the completion of this work.

## Appendix A. Proximal Algorithms

Let  $X = \mathbb{R}^N$ ,  $f \in \Gamma_0(X)$ ,  $g \in \Gamma_0(X)$ , using the tools from proximal calculus, we can solve the convex optimization problem with the general form

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{x}) + g(\mathbf{x}). \quad (\text{A.1})$$

Here, for simplicity, we assume each of the minimization problems, like (A.1) considered in this paper has a global minimizer. If the proximal operator of  $f + g$  was known or could be computed easily, we could recursively iterate the proximal operator to find a solution to (A.1). However, we often only know the proximal operator for  $f$  and  $g$  separately. In the following, we briefly introduce a few algorithms among the proximal algorithm category which can address this kind of minimization problem. Moreover, these algorithms can be adapted to be distributed across computing clusters [44, 47, 45, 19]; one example is used later in this work.

### Appendix A.1. Forward-Backward Splitting

In the case that  $f$  is differentiable, problem (A.1) can be solved using the Forward-Backward splitting algorithm. Starting with a proper initialization,  $\forall \lambda \in (0, +\infty)$ , the iterative scheme can be represented as

$$\mathbf{x}^{(k+1)} = \text{prox}_{\lambda g}(\mathbf{x}^{(k)} - \lambda \nabla f(\mathbf{x}^{(k)})), \quad (\text{A.2})$$

which includes a forward gradient step (explicit) regarding function  $f$  and a backward step (implicit and involves solving a proximal operator) with respect to  $g$ . Refer to [57, 58, 18, 17, 45, 19] and references therein for more details and the variants of the Forward-Backward splitting algorithm.

As an example, we see that formula (A.2) can be directly used to solve the unconstrained problem (3) so as to obtain an MAP estimator of the sky in radio astronomy. When  $g$  is the  $\ell_1$ -norm, this algorithm becomes the Iterative Shrinkage-Thresholding Algorithm (ISTA), where it is possible to obtain accelerated convergence by using Fast ISTA (FISTA) [58], which is detailed in Algorithm 4.

The Forward-Backward algorithm is often simpler to compute than the algorithms that follow, which is an advantage of solving the unconstrained problem over the constrained problem. However, there are many cases where  $f$  is not differentiable; for example when it represents an indicator function. Note that the Forward-Backward algorithm cannot be used to solve the constrained problem (17) directly, due to the non-differentiable indicator function.

**Algorithm 4** FISTA

---

```

1: given  $\mathbf{x}^{(0)} \in \mathbb{R}^N$ ,  $\lambda > 0$ ,  $\theta_0 = 1$ ,  $\hat{\mathbf{x}}^{(0)} = \mathbf{x}^{(0)}$ 
2: repeat for  $k = 0, \dots$ 
3:    $\mathbf{x}^{(k+1)} = \text{prox}_{\lambda g}(\hat{\mathbf{x}}^{(k)} - \lambda \nabla f(\hat{\mathbf{x}}^{(k)}))$ 
4:    $\theta_{k+1} = \frac{1 + \sqrt{1 + 4\theta_k^2}}{2}$ 
5:    $\hat{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k+1)} + \frac{\theta_k - 1}{\theta_{k+1}}(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})$ 
6: until convergence

```

---

*Appendix A.2. Douglas-Rachford Splitting*

When both  $f$  and  $g$  in (A.1) are non-differentiable, the Douglas-Rachford splitting algorithm can be applied; see [59, 60] for more details on the Douglas-Rachford splitting algorithm. Its iterative formula,  $\forall \lambda \in (0, +\infty)$ , reads

$$\begin{cases} \mathbf{x}^{(k)} = \text{prox}_{\lambda g}(\mathbf{v}^{(k)}), \\ \mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \gamma^{(k)}(\text{prox}_{\lambda f}(2\mathbf{x}^{(k)} - \mathbf{v}^{(k)}) - \mathbf{x}^{(k)}), \end{cases} \quad (\text{A.3})$$

where  $\gamma^{(k)} \in (\alpha, 2 - \alpha)$ ,  $\alpha \in (0, 1)$ . This iterative scheme needs the proximal operator for  $f$  and  $g$  individually. Therefore, the Douglas-Rachford splitting algorithm is restricted by the degree of difficulty of computing the proximal operators of  $f$  and  $g$ . The algorithm is summarized in Algorithm 5.

As an example, the Douglas-Rachford splitting algorithm can theoretically be used to solve the constrained problem (4) after moving its constraint into the objective functional by using the indicator function on an  $\ell_2$ -ball. However, if  $\Phi$  is not an identity operator, as in radio interferometry, solving the proximal operator of this kind of indicator function is not easy computationally.

**Algorithm 5** Douglas-Rachford Splitting Algorithm

---

```

1: given  $\mathbf{v}^{(0)} \in \mathbb{R}^N$ ,  $\alpha \in (0, 1)$ ,  $\lambda > 0$ 
2: repeat for  $k = 0, \dots$ 
3:    $\mathbf{x}^{(k)} = \text{prox}_{\lambda g}(\mathbf{v}^{(k)})$ 
4:    $\gamma^{(k)} \in (\alpha, 2 - \alpha)$ 
5:    $\mathbf{v}^{(k+1)} = \mathbf{v}^{(k)} + \gamma^{(k)}(\text{prox}_{\lambda f}(2\mathbf{x}^{(k)} - \mathbf{v}^{(k)}) - \mathbf{x}^{(k)})$ 
6: until convergence

```

---

*Appendix A.3. Alternating Direction Method of Multipliers*

The Forward-Backward and Douglas-Rachford splitting algorithms presented above require the proximal operators  $f$  and  $g$  to be easy to compute. In practice, this is sometimes not the case. For example, when function  $f$  involves explicitly a linear transformation  $\mathbf{L} \in \mathbb{R}^{K \times N}$  (e.g. a measurement operator), we must consider the problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{L}\mathbf{x}) + g(\mathbf{x}), \quad (\text{A.4})$$

where the proximal operator of  $f(\mathbf{L}\mathbf{x})$  has no explicit expression.

Problem (A.4) can be addressed by the alternating direction method of multipliers (ADMM) [47, 61, 19]. After setting  $\mathbf{v} = \mathbf{L}\mathbf{x}$ , problem (A.4) becomes

$$\min_{\mathbf{x} \in \mathbb{R}^N} f(\mathbf{v}) + g(\mathbf{x}), \quad \text{s.t.} \quad \mathbf{v} = \mathbf{L}\mathbf{x}. \quad (\text{A.5})$$

This problem has the following augmented Lagrangian with index  $\lambda \in (0, +\infty)$

$$\mathcal{L}(\mathbf{x}, \mathbf{v}, \mathbf{z}) := f(\mathbf{v}) + g(\mathbf{x}) + \frac{1}{\lambda} \mathbf{z}^\dagger (\mathbf{L}\mathbf{x} - \mathbf{v}) + \frac{1}{2\lambda} \|\mathbf{L}\mathbf{x} - \mathbf{v}\|_{\ell_2}^2, \quad (\text{A.6})$$

which can be solved alternatively corresponding to  $\mathbf{x}, \mathbf{v}, \mathbf{z}$ . More precisely,  $\mathcal{L}$  is minimized with respect to variables  $\mathbf{x}$  and  $\mathbf{v}$  alternatively while updating the dual variable  $\mathbf{z}$  (using the dual ascent method [47]) to ensure that the constraint  $\mathbf{v} = \mathbf{L}\mathbf{x}$  is met in the final solution, i.e.,

$$\mathbf{x}^{(k)} = \underset{\mathbf{x} \in \mathbb{R}^N}{\text{argmin}} \mathcal{L}(\mathbf{x}, \mathbf{v}^{(k)}, \mathbf{z}^{(k)}), \quad (\text{A.7})$$

$$\mathbf{v}^{(k+1)} = \underset{\mathbf{v} \in \mathbb{R}^K}{\text{argmin}} \mathcal{L}(\mathbf{x}^{(k)}, \mathbf{v}, \mathbf{z}^{(k)}), \quad (\text{A.8})$$

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + (\mathbf{L}\mathbf{x}^{(k)} - \mathbf{v}^{(k+1)}), \quad (\text{A.9})$$

which can be rewritten as

$$\mathbf{x}^{(k)} = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^N} \left( g(\mathbf{x}) + \frac{1}{2\lambda} \|\mathbf{L}\mathbf{x} - (\mathbf{v}^{(k)} - \mathbf{z}^{(k)})\|_{\ell_2}^2 \right), \quad (\text{A.10})$$

$$\mathbf{v}^{(k+1)} = \operatorname{argmin}_{\mathbf{v} \in \mathbb{R}^K} \left( f(\mathbf{v}) + \frac{1}{2\lambda} \|\mathbf{v} - (\mathbf{L}\mathbf{x}^{(k)} + \mathbf{z}^{(k)})\|_{\ell_2}^2 \right), \quad (\text{A.11})$$

$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + (\mathbf{L}\mathbf{x}^{(k)} - \mathbf{v}^{(k+1)}). \quad (\text{A.12})$$

Note, importantly, that the above problem (A.11) is actually computing the proximal operator of function  $f$  without involving the operator  $\mathbf{L}$ , which circumvents computing the proximal operator of  $f(\mathbf{L}\mathbf{x})$  directly and generally has an explicit expression. We comment that ADMM has a close relationship to the Douglas-Rachford algorithm, see [44, 45] for more details. The procedures of ADMM are briefly summarized in Algorithm 6, where we define

$$\operatorname{prox}_{\lambda g}^{\mathbf{L}}(\mathbf{u}) = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^N} \left( g(\mathbf{x}) + \frac{1}{2\lambda} \|\mathbf{L}\mathbf{x} - \mathbf{u}\|_{\ell_2}^2 \right), \quad (\text{A.13})$$

which may have a simple closed-form solution, or can be solved iteratively using a Forward-Backward method since its second term is differentiable.

---

**Algorithm 6** Alternating Direction Method of Multipliers (ADMM)

---

- 1: **given**  $\mathbf{z}^{(0)}, \mathbf{v}^{(0)} \in \mathbb{R}^K, \lambda > 0$
  - 2: **repeat** for  $k = 0, \dots$
  - 3:      $\mathbf{x}^{(k)} = \operatorname{prox}_{\lambda g}^{\mathbf{L}}(\mathbf{v}^{(k)} - \mathbf{z}^{(k)})$
  - 4:      $\mathbf{v}^{(k+1)} = \operatorname{prox}_{\lambda f}^{\mathbf{L}}(\mathbf{L}\mathbf{x}^{(k)} + \mathbf{z}^{(k)})$
  - 5:      $\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \mathbf{L}\mathbf{x}^{(k)} - \mathbf{v}^{(k+1)}$
  - 6: **until** convergence
- 

A generalization of ADMM is simultaneous direction method of multipliers (SDMM), which can be applied to an objective function of more than two functions [62, 15]. However, this method often requires operator inversion which can be expensive [19].

In this work, after setting  $f(\mathbf{L}\mathbf{x})$  to be the indicator function on an  $\ell_2$ -ball, we can use ADMM to solve the constrained problem (4) with the positivity constraint, i.e. the problem (17). We approximately solve  $\operatorname{prox}_{\lambda g}^{\mathbf{L}}(\mathbf{u})$  using an iteration of the Forward-Backward splitting method, and then use the Dual Forward-Backward splitting algorithm (which will be presented in Section Appendix A.5) to solve  $\operatorname{prox}_{\lambda g}(\mathbf{u})$  iteratively, where  $g$  contains the  $\ell_1$ -norm and the positivity constraint (see [19] and Section 5 for more detail).

#### Appendix A.4. Primal-Dual Splitting

In addition to ADMM, problem (A.4) can also be solved by the Primal-Dual splitting algorithm; an algorithm that like ADMM can be adapted to be distributed and performed in parallel [45, 19]. Firstly, the primal problem (A.4) can be rewritten as the following Primal-Dual formulation, i.e.,

$$\min_{\mathbf{x}} \max_{\mathbf{z}} g(\mathbf{x}) + \langle \mathbf{L}\mathbf{x}, \mathbf{z} \rangle - f^*(\mathbf{z}), \quad (\text{A.14})$$

which is a saddle point problem, where  $\langle \mathbf{L}\mathbf{x}, \mathbf{z} \rangle = \mathbf{z}^\dagger \mathbf{L}\mathbf{x}$ . It can be solved from minimizing and maximizing with respect to  $\mathbf{x}$  and  $\mathbf{z}$  alternatively, where for each subproblem the Forward-Backward ideas presented in Section Appendix A.1 can be applied if needed. The Primal-Dual algorithm is summarized in Algorithm 7. Furthermore, Moreau decomposition in equation (16) can be used to calculate the proximal operator of  $f^*$  given the proximal operator of  $f$ , i.e.

$$\operatorname{prox}_{\sigma f^*}(\mathbf{z}) = \mathbf{z} - \sigma \operatorname{prox}_{\sigma^{-1}f}(\mathbf{z}/\sigma).$$

The Primal-Dual and ADMM algorithms are both very efficient algorithms to solve problems like (A.4). The Primal-Dual algorithm generally can achieve better convergence rates than ADMM. However, since ADMM needs to compute the proximal operators  $\operatorname{prox}_{\lambda f}$  and  $\operatorname{prox}_{\lambda g}^{\mathbf{L}}$  and the Primal-Dual algorithm needs to compute  $\operatorname{prox}_{\sigma f^*}$  and  $\operatorname{prox}_{\tau g}$ , which method is more appropriate often depends on the overall problem itself. In addition, there has been plenty of work to optimize them further, which makes their performance more comparable and in some cases equivalent to each other (see [45] for an overview of Primal-Dual methods).

**Algorithm 7** Primal-Dual Algorithm

---

```

1: given  $\mathbf{x}^{(0)} \in \mathbb{R}^N, \mathbf{z}^{(0)} \in \mathbb{R}^N, \tau, \sigma > 0, \theta \in [0, 1]$ 
2: repeat for  $k = 0, \dots$ 
3:    $\mathbf{z}^{(k+1)} = \text{prox}_{\sigma f^*}(\mathbf{z}^{(k)} + \sigma \mathbf{L} \hat{\mathbf{x}}^{(k)})$ 
4:    $\mathbf{x}^{(k+1)} = \text{prox}_{\tau g}(\mathbf{x}^{(k)} - \tau \mathbf{L}^\dagger \mathbf{z}^{(k+1)})$ 
5:    $\hat{\mathbf{x}}^{(k+1)} = \mathbf{x}^{(k+1)} + \theta(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)})$ 
6: until convergence

```

---

*Appendix A.5. Dual Forward-Backward Splitting*

An algorithm closely related to the Primal-Dual algorithm is known as the Dual Forward-Backward splitting algorithm [63, 45]. To obtain the dual problem of (A.4), using Lagrangian multiplier  $\mathbf{z}$ , we get the Lagrangian

$$\mathcal{L}(\mathbf{x}, \mathbf{v}, \mathbf{z}) := f(\mathbf{v}) + g(\mathbf{x}) + \langle \mathbf{L}\mathbf{x} - \mathbf{v}, \mathbf{z} \rangle. \quad (\text{A.15})$$

By minimizing the Lagrangian over  $\mathbf{x}$  and  $\mathbf{v}$ , we have

$$\begin{aligned} \inf_{\mathbf{x}, \mathbf{v}} \mathcal{L}(\mathbf{x}, \mathbf{v}, \mathbf{z}) &= -\sup_{\mathbf{v}} (\langle \mathbf{z}, \mathbf{v} \rangle - f(\mathbf{v})) - \sup_{\mathbf{x}} (\langle -\mathbf{L}^\dagger \mathbf{z}, \mathbf{x} \rangle - g(\mathbf{x})) \\ &= -f^*(\mathbf{z}) - g^*(-\mathbf{L}^\dagger \mathbf{z}). \end{aligned} \quad (\text{A.16})$$

Then we have the dual problem of problem (A.4), i.e.

$$\min_{\mathbf{z}} f^*(\mathbf{z}) + g^*(-\mathbf{L}^\dagger \mathbf{z}). \quad (\text{A.17})$$

Note that term  $g^*(-\mathbf{L}^\dagger \mathbf{z})$  is differentiable and it is shown in [46] that

$$\partial_{\mathbf{z}} g^* = -\mathbf{L} \left( \underset{\mathbf{v} \in X}{\operatorname{argmin}} \left\{ \langle \mathbf{L}^\dagger \mathbf{z}, \mathbf{v} \rangle + g(\mathbf{v}) \right\} \right). \quad (\text{A.18})$$

Let  $\bar{g}(\mathbf{z}) = g^*(-\mathbf{L}^\dagger \mathbf{z})$ , applying the Forward-Backward splitting iterative scheme (A.2) with the relaxation on (A.17)  $f^* + g^* \rightarrow \sigma f^* + \sigma g^*$  with  $\sigma \in (0, \infty)$  and combining with the dual ascent method [47], we have

$$\mathbf{z}^{(k+1)} = \text{prox}_{\sigma f^*}(\mathbf{z}^{(k)} - \sigma \nabla \bar{g}(\hat{\mathbf{z}}^{(k)})), \quad (\text{A.19})$$

$$\hat{\mathbf{z}}^{(k+1)} = \mathbf{z}^{(k+1)} + \theta(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}). \quad (\text{A.20})$$

which is the so-call Dual Forward-Backward splitting algorithm.

In particular, applying the Forward-Backward splitting iterative scheme (A.2) on the minimization problem in (A.18), we have

$$\nabla \bar{g}(\mathbf{z}) = -\mathbf{L} \text{prox}_{\tau g}(\mathbf{x} - \tau \mathbf{L}^\dagger \mathbf{z}). \quad (\text{A.21})$$

Let  $\mathbf{x}^{(k+1)} = \text{prox}_{\tau g}(\mathbf{x}^{(k)} - \tau \mathbf{L}^\dagger \hat{\mathbf{z}}^{(k)})$  and substituting (A.21) into (A.19), we have the following iteration scheme

$$\mathbf{x}^{(k+1)} = \text{prox}_{\tau g}(\mathbf{x}^{(k)} - \tau \mathbf{L}^\dagger \hat{\mathbf{z}}^{(k)}) \quad (\text{A.22})$$

$$\mathbf{z}^{(k+1)} = \text{prox}_{\sigma f^*}(\mathbf{z}^{(k)} + \sigma \mathbf{L} \mathbf{x}^{(k+1)}) \quad (\text{A.23})$$

$$\hat{\mathbf{z}}^{(k+1)} = \mathbf{z}^{(k+1)} + \theta(\mathbf{z}^{(k+1)} - \mathbf{z}^{(k)}) \quad (\text{A.24})$$

After rearranging the order of the variables and replacing the relaxation strategy for  $\mathbf{z}$  by  $\mathbf{x}$ , the above Dual Forward-Backward splitting algorithm turns into the Primal-Dual algorithm (see Algorithm 7). See [45] for more discussions about the relation between the Dual Forward-Backward splitting algorithm and the Primal-Dual algorithm.

## References

- [1] L. L. McCready, J. L. Pawsey, R. Payne-Scott, Solar Radiation at Radio Frequencies and Its Relation to Sunspots, Proceedings of the Royal Society of London Series A 190 (1947) 357–375. doi:[10.1098/rspa.1947.0081](https://doi.org/10.1098/rspa.1947.0081).
- [2] M. Ryle, A. Hewish, The synthesis of large radio telescopes, MNRAS120 (1960) 220. doi:[10.1093/mnras/120.3.220](https://doi.org/10.1093/mnras/120.3.220).
- [3] R. Braun, T. Bourke, J. A. Green, E. Keane, J. Wagg, Advancing Astrophysics with the Square Kilometre Array, Advancing Astrophysics with the Square Kilometre Array (AASKA14) (2015) 174.
- [4] L. Koopmans, J. Pritchard, G. Mellesma, J. Aguirre, K. Ahn, R. Barkana, I. van Bemmel, G. Bernardi, A. Bonaldi, F. Briggs, A. G. de Bruyn, T. C. Chang, E. Chapman, X. Chen, B. Ciardi, P. Dayal, A. Ferrara, A. Fialkov, F. Fiore, K. Ichiki, I. T. Illiev, S. Inoue, V. Jelic, M. Jones, J. Lazio, U. Maio, S. Majumdar, K. J. Mack, A. Mesinger, M. F. Morales, A. Parsons, U. L. Pen, M. Santos, R. Schneider, B. Semelin, R. S. de Souza, R. Subrahmanyam, T. Takeuchi, H. Vedantham, J. Wagg, R. Webster, S. Wyithe, K. K. Datta, C. Trott, The Cosmic Dawn and Epoch of Reionisation with SKA, Advancing Astrophysics with the Square Kilometre Array (AASKA14) (2015) 1arXiv:1505.07568.
- [5] C. Ferrari, A. Dabbech, O. Smirnov, S. Makhathini, J. S. Kenyon, M. Murgia, F. Govoni, D. Mary, E. Slezak, F. Vazza, A. Bonafede, M. Brugger, M. Johnston-Hollitt, S. Dehghan, L. Feretti, G. Giovannini, V. Vacca, M. W. Wise, M. Gitti, M. Arnaud, G. Pratt, K. Zarb Adami, S. Colafrancesco, Non-thermal emission from galaxy clusters: feasibility study with SKA, Advancing Astrophysics with the Square Kilometre Array (AASKA14) (2015) 75arXiv:1412.5801.
- [6] M. Johnston-Hollitt, F. Govoni, R. Beck, S. Dehghan, L. Pratley, T. Akahori, G. Heald, I. Agudo, A. Bonafede, E. Carretti, T. Clarke, S. Colafrancesco, T. A. Ensslin, L. Feretti, B. Gaensler, M. Havercorn, S. A. Mao, N. Oppermann, L. Rudnick, A. Scaife, D. Schnitzeler, J. Stil, A. R. Taylor, V. Vacca, Using SKA Rotation Measures to Reveal the Mysteries of the Magnetised Universe, Advancing Astrophysics with the Square Kilometre Array (AASKA14) (2015) 92arXiv:1506.00808.
- [7] J. A. Högbom, Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines, A&AS15 (1974) 417.
- [8] U. Rau, T. J. Cornwell, A multi-scale multi-frequency deconvolution algorithm for synthesis imaging in radio interferometry, A&A532 (2011) A71. arXiv:1106.2745, doi:[10.1051/0004-6361/201117104](https://doi.org/10.1051/0004-6361/201117104).
- [9] L. Pratley, M. Johnston-Hollitt, An improved method for polarimetric image restoration in interferometry, MNRAS462 (2016) 3483–3501. arXiv:1606.01482, doi:[10.1093/mnras/stw1377](https://doi.org/10.1093/mnras/stw1377).
- [10] J. G. Ables, Maximum Entropy Spectral Analysis, A&AS15 (1974) 383.
- [11] T. J. Cornwell, K. F. Evans, A simple maximum entropy deconvolution algorithm, A&A143 (1985) 77–83.
- [12] Y. Wiaux, L. Jacques, G. Puy, A. M. M. Scaife, P. Vandergheynst, Compressed sensing imaging techniques for radio interferometry, MNRAS395 (2009) 1733–1742. arXiv:0812.4933, doi:[10.1111/j.1365-2966.2009.14665.x](https://doi.org/10.1111/j.1365-2966.2009.14665.x).
- [13] Y. Wiaux, G. Puy, Y. Boursier, P. Vandergheynst, Spread spectrum for imaging techniques in radio interferometry, MNRAS400 (2009) 1029–1038. arXiv:0907.0944, doi:[10.1111/j.1365-2966.2009.15519.x](https://doi.org/10.1111/j.1365-2966.2009.15519.x).
- [14] J. D. McEwen, Y. Wiaux, Compressed sensing for wide-field radio interferometric imaging, MNRAS413 (2011) 1318–1332. arXiv:1010.3658, doi:[10.1111/j.1365-2966.2011.18217.x](https://doi.org/10.1111/j.1365-2966.2011.18217.x).
- [15] R. E. Carrillo, J. D. McEwen, Y. Wiaux, PURIFY: a new approach to radio-interferometric imaging, MNRAS439 (2014) 3591–3604. arXiv:1307.4370, doi:[10.1093/mnras/stu202](https://doi.org/10.1093/mnras/stu202).
- [16] A. Dabbech, C. Ferrari, D. Mary, E. Slezak, O. Smirnov, J. S. Kenyon, MORESANE: MOdel REconstruction by Synthesis-ANalysis Estimators. A sparse deconvolution algorithm for radio interferometric imaging, A&A576 (2015) A7. arXiv:1412.5387, doi:[10.1051/0004-6361/201424602](https://doi.org/10.1051/0004-6361/201424602).
- [17] H. Garsden, J. N. Girard, J. L. Starck, S. Corbel, C. Tasse, A. Woiselle, J. P. McKean, A. S. van Amersfoort, J. Anderson, I. M. Avruch, R. Beck, M. J. Bentum, P. Best, F. Breitling, J. Broderick, M. Brüggen, H. R. Butcher, B. Ciardi, F. de Gasperin, E. de Geus, M. de Vos, S. Duscha, J. Eislöffel, D. Engels, H. Falcke, R. A. Fallows, R. Fender, C. Ferrari, W. Frieswijk, M. A. Garrett, J. Grießmeier, A. W. Gunst, T. E. Hassall, G. Heald, M. Hoeft, J. Hörandel, A. van der Horst, E. Juette, A. Karastergiou, V. I. Konratiev, M. Kramer, M. Kuniyoshi, G. Kuper, G. Mann, S. Markoff, R. McFadden, D. McKay-Bukowski, D. D. Mulcahy, H. Munk, M. J. Norden, E. Orru, H. Paas, M. Pandey-Pommier, V. N. Pandey, G. Pietka, R. Pizzo, A. G. Polatidis, A. Renting, H. Röttgering, A. Rowlinson, D. Schwarz, J. Sluman, O. Smirnov, B. W. Stappers, M. Steinmetz, A. Stewart, J. Swinbank, M. Tagger, Y. Tang, C. Tasse, S. Thoudam, C. Toribio, R. Vermeulen, C. Vocks, R. J. van Weeren, S. J. Wijnholds, M. W. Wise, O. Wucknitz, S. Yatawatta, P. Zarka, A. Zensus, LOFAR sparse image reconstruction, A&A575 (2015) A90. arXiv:1406.7242, doi:[10.1051/0004-6361/201424504](https://doi.org/10.1051/0004-6361/201424504).
- [18] J. Girard, H. Garsden, J. Starck, S. Corbel, A. Woiselle, C. Tasse, J. McKean, J. Bobin, Sparse representations and convex optimization as tools for lofar radio interferometric imaging, Journal of Instrumentation 10 (08) (2015) C08013. URL <http://stacks.iop.org/1748-0221/10/i=08/a=C08013>
- [19] A. Onose, R. E. Carrillo, A. Repetti, J. D. McEwen, J.-P. Thiran, J.-C. Pesquet, Y. Wiaux, Scalable splitting algorithms for big-data interferometric imaging in the SKA era, MNRAS462 (2016) 4314–4335. arXiv:1601.04026, doi:[10.1093/mnras/stw1859](https://doi.org/10.1093/mnras/stw1859).
- [20] L. Pratley, J. D. McEwen, M. d'Avezac, R. E. Carrillo, A. Onose, Y. Wiaux, Robust sparse image reconstruction of radio interferometric observations with purify, MNRAS473 (2018) 1038–1058. arXiv:1610.02400, doi:[10.1093/mnras/stx2237](https://doi.org/10.1093/mnras/stx2237).
- [21] L. Pratley, M. Johnston-Hollitt, J. D. McEwen, A fast and exact  $w$ -stacking and  $w$ -projection hybrid algorithm for wide-field interferometric imaging, ApJ, in press (Jul. 2019). arXiv:1807.09239.
- [22] A. Dabbech, A. Onose, A. Abdulaziz, R. A. Perley, O. M. Smirnov, Y. Wiaux, Cygnus A super-resolved via convex optimization from VLA data, MNRAS476 (2018) 2853–2866. arXiv:1710.08810, doi:[10.1093/mnras/sty372](https://doi.org/10.1093/mnras/sty372).
- [23] X. Cai, M. Pereyra, J. D. McEwen, Uncertainty quantification for radio interferometric imaging - I. Proximal MCMC methods, MNRAS480 (2018) 4154–4169. arXiv:1711.04818, doi:[10.1093/mnras/sty2004](https://doi.org/10.1093/mnras/sty2004).
- [24] X. Cai, M. Pereyra, J. D. McEwen, Uncertainty quantification for radio interferometric imaging: II. MAP estimation, MNRAS480 (2018) 4170–4182. arXiv:1711.04819, doi:[10.1093/mnras/sty2015](https://doi.org/10.1093/mnras/sty2015).
- [25] X. Cai, L. Pratley, J. D. McEwen, Online radio interferometric imaging: assimilating and discarding visibilities on arrival, MNRAS, in press (Dec. 2017). arXiv:1712.04462.
- [26] K. Akiyama, S. Ikeda, M. Pleau, V. L. Fish, F. Tazaki, K. Kuramochi, A. E. Broderick, J. Dexter, M. Mościbrodzka, M. Gowanlock, M. Honma, S. S. Doelean, Superresolution Full-polarimetric Imaging for Radio Interferometry with Sparse Modeling, AJ153 (2017) 159. arXiv:1702.00424, doi:[10.3847/1538-3881/aa6302](https://doi.org/10.3847/1538-3881/aa6302).
- [27] P. Broekema, R. van Nieuwpoort, H. Bal, The square kilometre array science data processor. preliminary compute platform design, Journal of Instrumentation 10 (07) (2015) C07004. URL <http://stacks.iop.org/1748-0221/10/i=07/a=C07004>
- [28] S. J. Tingay, R. Goeke, J. D. Bowman, D. Emrich, S. M. Ord, D. A. Mitchell, M. F. Morales, T. Boller, B. Crosse, R. B. Wayth, C. J. Lonsdale, S. Tremblay, D. Pallot, T. Colegate, A. Wicenec, N. Kudryavtseva, W. Arcus, D. Barnes, G. Bernardi, F. Briggs, S. Burns,

- J. D. Bunton, R. J. Cappallo, B. E. Corey, A. Deshpande, L. Desouza, B. M. Gaensler, L. J. Greenhill, P. J. Hall, B. J. Hazelton, D. Herne, J. N. Hewitt, M. Johnston-Hollitt, D. L. Kaplan, J. C. Kasper, B. B. Kincaid, R. Koenig, E. Kratzenberg, M. J. Lynch, B. McKinley, S. R. McWhirter, E. Morgan, D. Oberoi, J. Pathikulangara, T. Prabu, R. A. Remillard, A. E. E. Rogers, A. Roshi, J. E. Salah, R. J. Sault, N. Udaya-Shankar, F. Schlagenhaufer, K. S. Srivani, J. Stevens, R. Subrahmanyam, M. Waterson, R. L. Webster, A. R. Whitney, A. Williams, C. L. Williams, J. S. B. Wyithe, The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies, PASA30 (2013) 7. [arXiv:1206.6945](https://arxiv.org/abs/1206.6945), doi:10.1017/pasa.2012.007.
- [29] J. D. McEwen, A. M. M. Scaife, Simulating full-sky interferometric observations, MNRAS389 (2008) 1163–1178. [arXiv:0803.2165](https://arxiv.org/abs/0803.2165), doi:10.1111/j.1365-2966.2008.13690.x.
- [30] T. D. Carozzi, G. Woan, A generalized measurement equation and van Cittert-Zernike theorem for wide-field radio astronomical interferometry, MNRAS395 (2009) 1558–1568. [arXiv:0812.0141](https://arxiv.org/abs/0812.0141), doi:10.1111/j.1365-2966.2009.14642.x.
- [31] O. M. Smirnov, Revisiting the radio interferometer measurement equation. IV. A generalized tensor formalism, A&A531 (2011) A159. [arXiv:1106.0579](https://arxiv.org/abs/1106.0579), doi:10.1051/0004-6361/201116764.
- [32] D. C. Price, O. M. Smirnov, Generalized formalisms of the radio interferometer measurement equation, MNRAS449 (2015) 107–118. [arXiv:1501.06447](https://arxiv.org/abs/1501.06447), doi:10.1093/mnras/stv137.
- [33] F. Zernike, The concept of degree of coherence and its application to optical problems, Physica 5 (1938) 785–795. doi:10.1016/S0031-8914(38)80203-2.
- [34] J. A. Fessler, B. P. Sutton, Nonuniform fast fourier transforms using min-max interpolation, IEEE Transactions on Signal Processing 51 (2003) 560–574. doi:10.1109/TSP.2002.807005.
- [35] A. R. Thompson, J. Moran, G. Swenson, Interferometry and Synthesis in Radio Astronomy, Wiley, 2008.
- [36] M. Pereyra, J. M. Bioucas-Dias, M. A. T. Figueiredo, Maximum-a-posteriori estimation with unknown regularisation parameters, in: 2015 23rd European Signal Processing Conference (EUSIPCO), 2015, pp. 230–234. doi:10.1109/EUSIPCO.2015.7362379.
- [37] R. E. Carrillo, J. D. McEwen, Y. Wiaux, Sparsity Averaging Reweighted Analysis (SARA): a novel algorithm for radio-interferometric imaging, MNRAS426 (2012) 1223–1234. [arXiv:1205.3123](https://arxiv.org/abs/1205.3123), doi:10.1111/j.1365-2966.2012.21605.x.
- [38] M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors, in: 2006 14th European Signal Processing Conference, 2006, pp. 1–5.
- [39] J. Birdi, A. Repetti, Y. Wiaux, Sparse interferometric Stokes imaging under the polarization constraint (Polarized SARA), MNRAS478 (2018) 4442–4463. [arXiv:1801.02417](https://arxiv.org/abs/1801.02417), doi:10.1093/mnras/sty1182.
- [40] J. Birdi, A. Repetti, Y. Wiaux, Scalable algorithm for polarization constrained sparse interferometric stokes imaging, in: 2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM), 2018, pp. 465–469. doi:10.1109/SAM.2018.8448968.
- [41] A. Abdulaziz, A. Dabbech, A. Onose, Y. Wiaux, A low-rank and joint-sparsity model for hyper-spectral radio-interferometric imaging, in: 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 388–392. doi:10.1109/EUSIPCO.2016.7760276.
- [42] J. Deguignet, A. Ferrari, D. Mary, C. Ferrari, Distributed multi-frequency image reconstruction for radio-interferometry, in: 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1483–1487. doi:10.1109/EUSIPCO.2016.7760495.
- [43] S. Boyd, S. Boyd, L. Vandenberghe, C. U. Press, **Convex Optimization**, Berichte über verteilte messsysteme, Cambridge University Press, 2004.  
URL <https://books.google.co.uk/books?id=mYm0bLd3fcoC>
- [44] P. L. Combettes, J.-C. Pesquet, **Proximal Splitting Methods in Signal Processing**, Springer New York, New York, NY, 2011, pp. 185–212. doi:10.1007/978-1-4419-9569-8\_10.  
URL [https://doi.org/10.1007/978-1-4419-9569-8\\_10](https://doi.org/10.1007/978-1-4419-9569-8_10)
- [45] N. Komodakis, J. Pesquet, Playing with duality: An overview of recent primal?dual approaches for solving large-scale optimization problems, IEEE Signal Processing Magazine 32 (6) (2015) 31–54. doi:10.1109/MSP.2014.2377273.
- [46] P. Giselsson, S. Boyd, Metric selection in fast dual forwardâ\$backward splitting, Automatica 62 (2015) 1 – 10. doi:<https://doi.org/10.1016/j.automatica.2015.09.010>.  
URL <http://www.sciencedirect.com/science/article/pii/S0005109815003611>
- [47] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Found. Trends Mach. Learn. 3 (1) (2011) 1–122. doi:10.1561/2200000016.  
URL <http://dx.doi.org/10.1561/2200000016>
- [48] M. Snir, W. Gropp, S. Otto, S. Huss-Lederman, J. Dongarra, D. Walker, **MPI—the Complete Reference: The MPI core**, MPI.: The Complete Reference : The MPI Core, Mass, 1998.  
URL <https://books.google.co.uk/books?id=x79puJ2YkroC>
- [49] L. Pratley, J. D. McEwen, M. d’Avezac, R. Carrillo, I. Christidi, R. Guichard, D. Pérez-Suárez, Y. Wiaux, **PURIFY** (Feb. 2019). doi:10.5281/zenodo.2587838.  
URL <https://doi.org/10.5281/zenodo.2587838>
- [50] L. Pratley, J. D. McEwen, M. d’Avezac, R. Carrillo, I. Christidi, R. Guichard, D. Pérez-Suárez, Y. Wiaux, **SOPT** (Feb. 2019). doi:10.5281/zenodo.2584256.  
URL <https://doi.org/10.5281/zenodo.2584256>
- [51] G. Guennebaud, B. Jacob, et al., Eigen v3, <http://eigen.tuxfamily.org> (2010).
- [52] I. Daubechies, W. Sweldens, Factoring wavelet transforms into lifting steps, Journal of Fourier analysis and applications 4 (3) (1998) 247–269.
- [53] S. Mallat, A wavelet tour of signal processing, Elsevier, 1999.
- [54] P. Yalamanchili, U. Arshad, Z. Mohammed, P. Garigipati, P. Entschev, B. Kloppenborg, J. Malcolm, J. Melonakos, **ArrayFire - A high performance software library for parallel computing with an easy-to-use API** (2015).  
URL <https://github.com/arrayfire/arrayfire>
- [55] E. W. Greisen, **AIPS FITS File Format (revised)**, AIPS MEMORANDUM 117, National Radio Astronomy Observatory, Charlottesville, Virginia (2016).  
URL <http://www.aips.nrao.edu/aipsmemo.html>
- [56] P. J. Beatty, D. G. Nishimura, J. M. Pauly, Rapid gridding reconstruction with a minimal oversampling ratio, IEEE Transactions on Medical Imaging 24 (2005) 799–808.
- [57] I. Daubechies, M. Defrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, Communications on Pure and Applied Mathematics 57 (11) (2004) 1413–1457. [arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20042](https://arxiv.org/abs/https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20042), doi:10.1002/cpa.20042.  
URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20042>
- [58] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, SIAM Journal on Imaging Sciences 24

- 2 (1) (2009) 183–202. [arXiv:<https://doi.org/10.1137/080716542>](https://doi.org/10.1137/080716542), doi:[10.1137/080716542](https://doi.org/10.1137/080716542).
- URL <https://doi.org/10.1137/080716542>
- [59] P. L. Combettes, J. Pesquet, A douglas–rachford splitting approach to nonsmooth convex variational signal recovery, IEEE Journal of Selected Topics in Signal Processing 1 (4) (2007) 564–574. doi:[10.1109/JSTSP.2007.910264](https://doi.org/10.1109/JSTSP.2007.910264).
- [60] R. Bot, C. Hendrich, A douglas–rachford type primal-dual method for solving inclusions with mixtures of composite and parallel-sum type monotone operators, SIAM Journal on Optimization 23 (4) (2013) 2541–2565. [arXiv:<https://doi.org/10.1137/120901106>](https://doi.org/10.1137/120901106), doi:[10.1137/120901106](https://doi.org/10.1137/120901106).
- URL <https://doi.org/10.1137/120901106>
- [61] J. Yang, Y. Zhang, Alternating direction algorithms for l1-problems in compressive sensing, SIAM Journal on Scientific Computing 33 (1) (2011) 250–278. [arXiv:<https://doi.org/10.1137/090777761>](https://doi.org/10.1137/090777761), doi:[10.1137/090777761](https://doi.org/10.1137/090777761).
- URL <https://doi.org/10.1137/090777761>
- [62] S. Setzer, G. Steidl, T. Teuber, Deblurring poissonian images by split bregman techniques, Journal of Visual Communication and Image Representation 21 (3) (2010) 193 – 199. doi:<https://doi.org/10.1016/j.jvcir.2009.10.006>.
- URL <http://www.sciencedirect.com/science/article/pii/S1047320309001539>
- [63] P. L. Combettes, D. Dũng, B. C. Vũ, Dualization of signal recovery problems, Set-Valued and Variational Analysis 18 (3) (2010) 373–404. doi:[10.1007/s11228-010-0147-7](https://doi.org/10.1007/s11228-010-0147-7).
- URL <https://doi.org/10.1007/s11228-010-0147-7>