Nr. 68
10. February 2020

universität innsbruck
Institut für Mathematik

Preprint-Series: Department of Mathematics - Applied Mathematics

Unsupervised adaptive neural network regularization for accelerated radial cine MRI

A. Kofler, M. Dewey, T. Schaeffter, C. Kolbitsch, M. Haltmeier

APPLIEDMATHEMATICS

Technikerstraße 13 - 6020 Innsbruck - Austria
Tel.: +43 512 507 53803 Fax: +43 512 507 53898
https://applied-math.uibk.ac.at

Unsupervised Adaptive Neural Network Regularization for Accelerated Radial Cine MRI

Andreas Kofler, Marc Dewey, Tobias Schaeffter, Christoph Kolbitsch and Markus Haltmeier

***Abstract*—In this work, we propose an iterative reconstruction scheme (ALONE - Adaptive Learning Of NEtworks) for 2D radial cine MRI based on ground truth-free unsupervised learning of shallow convolutional neural networks. The network is trained to approximate patches of the current estimate of the solution during the reconstruction. By imposing a shallow network topology and constraining the L_2-norm of the learned filters, the network's representation power is limited in order not to be able to recover noise. Therefore, the network can be interpreted to perform a low dimensional approximation of the patches for stabilizing the inversion process. We compare the proposed reconstruction scheme to two ground truth-free reconstruction methods, namely a well known Total Variation (TV) minimization and an unsupervised adaptive Dictionary Learning (DIC) method. The proposed method outperforms both methods with respect to all reported quantitative measures. Further, in contrast to DIC, where the sparse approximation of the patches involves the solution of a complex optimization problem, ALONE only requires a forward pass of all patches through the shallow network and therefore significantly accelerates the reconstruction.**

***Index Terms*—Neural Networks, Unsupervised Learning, Inverse Problems, Dynamic MRI, Image Processing, Compressed Sensing, Iterative Reconstruction**

I. Introduction

Magnetic Resonance Imaging (MRI) is a widely used and indispensable medical tool for the non-invasive assessment of various diseases. For example, dynamic cardiac MRI, allows for the assessment of the cardiac function. Thereby, a certain number of cardiac phases is obtained.

However, MRI is well-known to suffer from relatively long acquisition times which for example limit the achievable temporal resolution which is required for a proper diagnosis. Therefore, in order to accelerate the measurement process, different techniques have emerged in field of MRI. For example, Parallel Imaging [1], [2] allows the acceleration of the data-acquisition process as solution implemented on a hardware level. In addition, to further accelerate the acquisition process, instead of acquiring the full k-space data in Fourier-domain, undersampling schemes have been extensively investigated in the literature.

Compressed Sensing theory [3], [4] delivers theoretical guarantees and error bounds on the the signal-recovery from a set of random measurements. However, randomly sampling k-space data is challenging from a technical point of view and therefore, different undersampling schemes have been investigated in the literature, [5], [6]. In particular, radial undersampling has the advantages of oversampling the center of k-space which contains the Fourier-coefficients corresponding to the basis functions with lower frequencies. As most physiological motions are smooth (e.g. heart contraction), radial undersampling has been reported to be particularly suitable for dynamic cine MRI applications. When undersampling k-space data, the underlying inverse problem becomes an underdetermined one and no unique solution exists. Therefore, several regularization techniques have been proposed in the past, either based on hand-crafted priors, e.g. Total Variation (TV)-minimization [7], or regularizations based on information learned from data, e.g. Dictionary Learning [8], [9], [10].

Recently, Neural Networks (NNs) have been widely applied within the field of inverse problems. Most commonly, the convolutional NNs (CNNs) are either applied as post-processing methods to reduce artefacts or denoise images, see for example [11], [12], [13], [14] or employed in so-called iterative or cascaded neural networks [15], [16], [17], [18], [19]. In the latter, the network architectures consist of CNNs as well as layers containing the forward and the adjoint operators which are used to ensure that the output of the CNNs match the acquired raw data. However, most methods based on CNNs nowadays are based on supervised learning (SL), i.e. on the implicit assumption of the availability of a large enough dataset of pairs. The literature in which CNNs using Unsupervised Learning (UL) are applied, is highly under-represented. In this work, we propose a method for image reconstruction in undersampled $2D$ radial cine MRI using an adaptive unsupervised learning approach, where the regularization is learned during the reconstruction process. Let $\mathbf{A}_I \colon \mathbb{C}^N \to \mathbb{C}^m$ be the undersampled dynamic radial cine MRI forward operator, $\mathbf{y}_I \simeq \mathbf{A}_I \mathbf{x}$ the available k-space data of the unknown image $\mathbf{x}$. Let $\mathbf{E}_j \colon \mathbb{C}^N \to \mathbb{C}^d$ denote the operator that extracts the j-th patch and $\mathbf{\Phi}_\theta \colon \mathbb{C}^d \to \mathbb{C}^d$ a neural network (see Section II-A for precise formulations).

A. Kofler is with the Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany (e-mail: andreas.kofler@charite.de)

M. Dewey is with the Department of Radiology, Charité - Universitätsmedizin Berlin, Berlin, Germany and the Berlin Institute of Health, Berlin, Germany (e-mail: marc.dewey@charite.de)

T. Schaeffter is with the Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany, Kings College London, London, UK and the Department of Medical Engineering, Technical University of Berlin, Berlin, Germany (e-mail: tobias.schaeffter@ptb.de)

C. Kolbitsch is with the Physikalisch-Technische Bundesanstalt (PTB), Braunschweig and Berlin, Germany and Kings College London, London, UK (e-mail: christoph.kolbitsch@ptb.de)

M. Haltmeier is with the Department of Mathematics, University of Innsbruck, Innsbruck, Austria (e-mail:markus.haltmeier@uibk.ac.at)

The proposed approach is based on minimizing the functional

$$\mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x},\theta) \triangleq \frac{1}{2}\|\mathbf{A}_I\mathbf{x}-\mathbf{y}_I\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\|\mathbf{E}_j(\mathbf{x})-\mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)\|_2^2 + \Omega(\theta) \quad (1)$$

jointly over $\mathbf{x} \in X$ and the set of trainable parameters $\theta \in \mathbb{R}^q$. The term $\sum_{j=1}^{p}\|\mathbf{E}_j(\mathbf{x})-\mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)\|_2^2$ acts as regularizer defined by a neural network that is adapted to the specific data, i.e. all the available image-patches, and Ω is a penalty that prevents overfitting of the network to noise.

In order to minimize (1), we propose an iterative minimization procedure (ALONE- Adaptive Learning Of NEtworks) that performs minimization steps in $\mathbf{x}$ and θ in an alternating manner. The update step of the network parameters amounts to network training on patches of the current iterates. Therefore the network is trained in a completely unsupervised manner without needing to rely on ground truth image data. As we shall demonstrate, ALONE can be used with rather small patch size and shallow convolutional neural networks. As a result, ALONE is numerically efficient, comes without any preceding training phase and does not require artefact-free ground truth data. To the best of our knowledge, we are not aware of any CNN-based image reconstruction method sharing similar features.

The rest of the paper is structured as follows. In Section II, the inverse problem as well the proposed reconstruction algorithm are formally introduced and discussed. In Section III, we introduce the quantitative measures which are used to evaluate the performance of our method. We compare it to the well-known TV-minimization approach, a dictionary learning-based approach and a method using previously trained CNNs to generate priors which are then used in an iterative reconstruction. We then conclude the work with a discussion and some conclusions in Section IV and V. Note that while we focus our presentation on dynamic radial MRI, we point out that the proposed framework can be used for general 2D or 3D image reconstruction problems as well.

II. Proposed Reconstruction Framework

In this section, we give a precise problem formulation and introduce the proposed unsupervised adaptive deep neural network based reconstruction framework.

A. Problem Formulation

We consider the problem given by

$$\mathbf{A}_I\mathbf{x} = \mathbf{y}_I, \quad (2)$$

where the forward operator $\mathbf{A}_I$ is given by the composition $\mathbf{S}_I \circ \mathbf{A}$, of a binary mask $\mathbf{S}_I$ and the $\mathbf{A}$ the (discretized) $2D$ frame-wise Fourier-encoding operator which samples the k-space data along radial lines. More precisely, the radial trajectories are chosen according to the golden-angle radial method [20], [21]. The coefficients are assumed to be enumerated by a set of indices $I \subset J = \{1,\dots,N_{\mathrm{rad}}\}$ with $|I| \triangleq m < N_{\mathrm{rad}}$ which corresponds to a subset of all N_{rad} Fourier coefficients that could be sampled. The number N_{rad} is more precisely specified by the MR-acquisition parameters, i.e. by the number of radial trajectories, the number of receiver coils, the number of acquired cardiac phases, etc. For further details about a possible implementation of the radial Fourier-encoding operator, we refer to [22].

The vector $\mathbf{y}_I \in \mathbb{C}^m$ contains the undersampled k-space data and the goal is to reconstruct a complex-valued 3D image $\mathbf{x} \in \mathbb{C}^N$ with $N = N_x \times N_y \times N_t$ from the measurements $\mathbf{y}_I$. Due to the application of the binary mask $\mathbf{S}_I$, the Nyquist criterion is violated and the direct reconstruction from the measured data yields images which are contaminated by artefacts . Addressing the undersampling issue requires the use of proper regularization techniques which exploit structure in the manifolds of potential solutions to provide high quality and aliasing artefact-free results.

In the following, for convenience, we write

$$\big\|\mathbf{E}(\mathbf{x})-\mathbf{\Phi}_\theta\big(\mathbf{E}(\mathbf{x})\big)\big\|_2^2 \triangleq \sum_{j=1}^{p}\|\mathbf{E}_j(\mathbf{x})-\mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)\|_2^2$$
$$\mathbf{E}(\mathbf{x}) \triangleq \big(\mathbf{E}_1(\mathbf{x}),\dots,\mathbf{E}_p(\mathbf{x})\big)$$
$$\mathbf{\Phi}_\theta\big(\mathbf{E}(\mathbf{x})\big) \triangleq \Big(\mathbf{\Phi}_\theta\big(\mathbf{E}_1(\mathbf{x})\big),\dots,\mathbf{\Phi}_\theta\big(\mathbf{E}_p(\mathbf{x})\big)\Big).$$

Here, $\mathbf{E}_j\colon \mathbb{C}^N \to \mathbb{C}^d$ for $j = 1,\dots,p$ is the operator which extracts the j-th 3D patch (i.e. a small sub-portion of the image) from an image $\mathbf{x}$, $\mathbf{\Phi}_\theta\colon \mathbb{C}^d \to \mathbb{C}^d$ is a neural network with trainable parameters $\theta \in \mathbb{R}^q$ operating on the patches. The number p results from the shape of the patches and the strides used to extract the patches. As presented in the introduction, we approach the reconstruction problem as finding a regularized solution $\mathbf{x} \in \mathbb{C}^N$ by jointly minimizing (1) over $\mathbf{x} \in \mathbb{C}^N$ and $\theta \in \mathbb{R}^q$. Using the above introduced notation, this amounts to the optimization problem

$$\mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x},\theta) = \frac{1}{2}\|\mathbf{A}_I\mathbf{x}-\mathbf{y}_I\|_2^2 + \frac{\lambda}{2}\big\|\mathbf{E}(\mathbf{x})-\mathbf{\Phi}_\theta\big(\mathbf{E}(\mathbf{x})\big)\big\|_2^2 + \Omega(\theta) \to \min_{\mathbf{x},\theta}. \quad (3)$$

Here, Ω denotes a regularization imposed on the parameter set $\mathbb{R}^q$. It limits the capacity of the network $\mathbf{\Phi}_\theta$ such that it does not adapt to image noise. In order to minimize (1) we propose an alternating minimization algorithm described below.

B. Proposed Reconstruction Algorithm

We begin the reconstruction process by applying the adjoint operator to the measured data and obtaining an initial guess of the solution $\mathbf{x}_I \triangleq \mathbf{A}_I^{\mathsf{H}}\mathbf{y}_I$. Then, we proceed by alternating between the following minimization steps (R1) and (R2) with respect to $\mathbf{x}$ and θ, respectively.

(R1) **Network update:** We first update the set of parameters $\theta \in \mathbb{R}^q$. For this purpose, we fix $\mathbf{x} \in \mathbb{C}^N$ and solve

$$\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta) \triangleq \frac{\lambda}{2}\sum_{j=1}^{p} \|\mathbf{E}_j(\mathbf{x}) - \mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)\|_2^2 + \Omega(\theta) \to \min_\theta . \quad (4)$$

Minimizing the loss function $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta)$ clearly corresponds to training the network $\mathbf{\Phi}_\theta$ on a dataset of pairs of patches which are extracted from the current image estimate $\mathbf{x}$. The aim of the network is to reproduce the relevant (low-dimensional) information contained in the patches $\mathbf{E}_j(\mathbf{x})$ and discard the (high-dimensional) noise-like artefacts. Therefore, for each patch, the network can be interpreted to perform a low-dimensional approximation of the patches. In practice, problem (4) can be efficiently solved by employing state-of-the-art non-linear optimization routines, e.g, the ADAM optimizer [23].

(R2) **Reconstruction update:** After having obtained an estimate for θ, we set $\mathbf{z}_j = \mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big) \in \mathbb{C}^d$ for any patch and and update the image estimate $\mathbf{x} \in \mathbb{C}^N$ by solving

$$\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x}) \triangleq \frac{1}{2}\|\mathbf{A}_I\mathbf{x} - \mathbf{y}_I\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{p} \|\mathbf{E}_j(\mathbf{x}) - \mathbf{z}_j\|_2^2 \to \min_{\mathbf{x}} . \quad (5)$$

The optimization problem (5) is quadratic and hence can be solved efficiently. More precisely, according to Fermat's rule, $\mathbf{x}$ solves (5) if and only if it satisfies the linear optimality condition

$$\mathbf{0} = \nabla_{\mathbf{x}}\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x}) = \mathbf{H}\mathbf{x} - \mathbf{c}\,, \quad (6)$$

with

$$\mathbf{H} \triangleq \mathbf{A}_I^{\mathsf{H}}\mathbf{A}_I + \lambda\sum_{j=1}^{p}\mathbf{E}_j^{\mathsf{T}}\mathbf{E}_j, \quad (7)$$

$$\mathbf{c} \triangleq \mathbf{x}_I + \lambda\sum_{j=1}^{p}\mathbf{E}_j^{\mathsf{T}}(\mathbf{z}_j). \quad (8)$$

If the operator $\mathbf{A}$ is an isometry, e.g. when the full data acquisition takes place using a single-coil and sampling along a Cartesian grid, it holds $\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{x}\|_2$ for all $\mathbf{x}$ and, consequently, problem (5) has an analytic solution. It is given by performing a linear combination of the available k-space data $\mathbf{y}_I$ and the one estimated by CNN-approximation and then subsequently applying the inverse operator, i.e.

$$\mathbf{x}^* = \mathbf{A}^{\mathsf{H}}\Big(\frac{\lambda}{1+\lambda}\mathbf{y}_I + \mathbf{\Lambda}\mathbf{A}\sum_{j=1}^{p}\mathbf{E}_j^{\mathsf{T}}(\mathbf{z}_j)\Big), \quad (9)$$

where the diagonal operator $\mathbf{\Lambda}$ accounts for proper weighting of the k-space data; see [8] for a detailed derivation of (9). In the general case, the solution of problem (5) can be obtained by solving the linear matrix equation $\mathbf{H}\mathbf{x} = \mathbf{c}$. The system $\mathbf{H}\mathbf{x} = \mathbf{c}$ can be efficiently solved by means of any iterative scheme. Due to the symmetric structure of the operator $\mathbf{H}$, we can apply, for example, the pre-conditioned conjugate gradient method (PCG) [24].

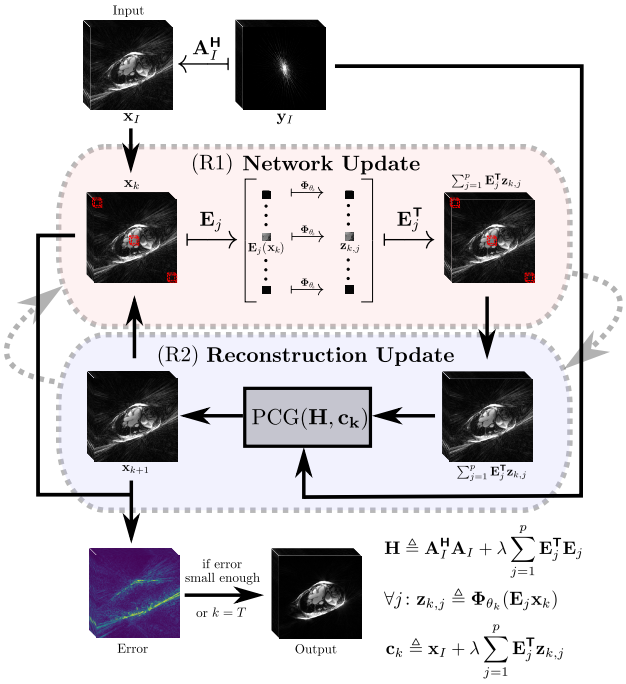

Figure 1. Reconstruction algorithm ALONE. Network update: from the current image estimate $\mathbf{x}_k$, 3D patches are extracted and used for training the CNN $\mathbf{\Phi}_\theta$ in an unsupervised manner. Then, after training, all patches are processed using the CNN $\mathbf{\Phi}_\theta$ and reassembled to obtain a regularized solution $\mathbf{\Phi}_\theta(\mathbf{x}_k)$. The regularized solution is then used in a reconstruction update step, which updates the image estimate using PCG.

Algorithm 1 Proposed ALONE algorithm

Input: Initialization $\mathbf{x}_0 = \mathbf{A}_I^{\mathsf{H}}\mathbf{y}_I$
Parameters: $\lambda > 0$, iteration number $T > 0$, accuracy $\varepsilon \geq 0$
Output: reconstructed image $\mathbf{x}_{\text{reco}}$

$k \leftarrow 0$
$e_k \leftarrow \infty$
while $k \leq T$ and $e_k > \varepsilon$ **do**
  $\theta_k \leftarrow \arg\min_\theta \mathcal{L}_{\mathbf{x}_k,\mathbf{y}_I,\lambda}(\theta)$
  $\forall j\colon \mathbf{z}_{k,j} \leftarrow \mathbf{\Phi}_{\theta_k}\big(\mathbf{E}_j(\mathbf{x}_k)\big)$
  $\mathbf{c}_k \leftarrow \mathbf{x}_I + \lambda\sum_{j=1}^{p}\mathbf{E}_j^{\mathsf{T}}(\mathbf{z}_{k,j})$
  $\mathbf{x}_{k+1} \leftarrow \arg\min_{\mathbf{x}} \mathcal{L}_{\theta_k,\mathbf{y}_I,\lambda}(\mathbf{x})$ by solving $\mathbf{H}\mathbf{x} = \mathbf{c}_k$
  $e_{k+1} \leftarrow \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2/\|\mathbf{x}_k\|_2^2$
  $k \leftarrow k + 1$
end while
$\mathbf{x}_{\text{reco}} \leftarrow x_k$

After having obtained the solutions to problem (4) and (5), we repeat the procedure until a pre-defined stopping criterion is fulfilled. Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be the sequence of reconstructions obtained as just described. We stop the iteration either if the relative change of the newly obtained solution $\mathbf{x}_{k+1}$ is small enough, i.e. $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2/\|\mathbf{x}_k\|_2^2 < \varepsilon$ for some $\varepsilon \geq 0$ or if a chosen maximal number of iterations $T > 0$ has been performed. Algorithm 1 summarizes the just described steps, which we name ALONE (Adaptive Learning Of NEtworks) reconstruction algorithm. Figure 1 shows an illustration of the Algorithm.

III. Experiments

A. Dataset and Evaluation Metrics

For the evaluation of our proposed method, we used a dataset of four patients for which we acquired $2D$ cine MRI image sequences. The image sequences have an in-plane number of pixels of $N_x \times N_y = 320 \times 320$ and a number of cardiac phases of $N_t = 30$. For two of the patients, $N_z = 12$ different slices were obtained, while for the resting two patients, we could only acquire $N_z = 6$ slices due to restricted respiratory capabilities. Thus, our dataset consists of a set of 36 two-dimensional cine MR image sequences. In order to quantitatively assess the performance of our method, we reconstructed all the $2D$ cine MR images using kt-SENSE [25], [21] using $N_\varphi = 3400$ radial lines. Then, from the obtained image sequences we retrospectively generated radially acquired k-space data by only sampling along $N_\varphi = 1130$ radial trajectories. More precisely, in our case, the operator $\mathbf{A}_I$ is given by $\mathbf{A}_I = \operatorname{diag}(\mathbf{S}_I, \dots, \mathbf{S}_I) \circ \operatorname{diag}(\mathbf{A}, \dots, \mathbf{A}) \circ [\mathbf{C}_1, \dots, \mathbf{C}_{n_c}]^\mathsf{T}$, where $\mathbf{C}_i$ is the i-th coil-sensitivity map and $n_c = 12$ and, again, $\mathbf{A}$ is the frame-wise radial Fourier encoding operator. Note that due to the used radial sampling pattern, sampling along $N_\varphi = 3400$ spokes already corresponds to an undersampling factor of ~ 3. Therefore, acquiring k-space data along only $N_\varphi = 1130$ corresponds to an acceleration factor of ~ 9.
We assessed the quality of our obtained reconstructions by comparing them to the kt-SENSE reconstructions obtained using $N_\varphi = 3400$ radial spokes. For the evaluation, we used the following quantitative measures: peak signal-to-noise ratio (PSNR), normalized root mean squared error (NRMSE), the structural similarity index measure (SSIM) and the Haar wavelet-based perceptual similarity measure (HPSI) [26]. Since the field of view is quite large and image sequences contain a noticeable portion of background which is irrelevant for diagnostic purposes, before calculating the statistics, we cropped all the image sequences to $N_x \times N_y \times N_t = 160 \times 160 \times 30$ using a symmetric cut-off of 80 in x- and y-direction.

B. Network Architecture and Training

Here, we briefly describe the network architecture used for all the experiments. The CNN is shown in Figure 2. It consists of a three-layers CNN with only one hidden layer. The input of the CNN is a patch $\mathbf{E}_j(\mathbf{x})$ which is extracted from the current estimate of the image. Since the images are complex-valued we represent the patches using two-channels. The image patch is passed through a $3 \times 3 \times 3$-convolutional layer with K filters, followed by a voxel-wise application of the ReLU activation function. Then, from the feature maps a complex-valued patch is obtained by applying a $1 \times 1 \times 1$-convolutional layer with the identity as activation function. Therefore, the output patch corresponds to a learned linear combination of the extracted K feature maps which are learned by the K filters.
Intuitively speaking, the network $\mathbf{\Phi}_\theta$ is trained to perform a learned dimensionality reduction of the $3D$ patches which are supposed to lie on a lower dimensional manifold. Similar to dictionary learning, where signals are represented as sparse combinations of elements of an overcomplete basis, our method performs a dimensionality reduction representing each $3D$ patch as a linear combination of last extracted feature maps which depend on the learned K filters. However, in contrast to dictionary learning, where, once the dictionary is learned, the correspondent support of the signals has to be calculated by some sparse coding algorithm, our method extracts K filters which can be globally used for all the patches.
Since the network $\mathbf{\Phi}_\theta$ is trained in an unsupervised manner on the patches of the current image estimate, the network's representation power has to be constrained in order make it a proper regularization. The first restriction is directly given by the fact that the network is very shallow and only contains one hidden layer. Second, the number of learned filters is chosen to be quite small, for example $K = 16$. Further, while training the network, a further regularization $\Omega(\theta)$ is included in the loss function. We choose to bound the L_2-norm of the learned kernels, i.e. $\Omega(\theta) = \sum_{k=1}^K \|f_k\|_2^2$, where f_k is the k-th convolutional filter.

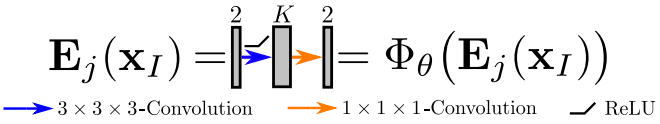

Figure 2. The shallow network used in the experiments. The input is a complex-valued $3D$ patch which is extracted from the image sequence. Since the data is complex-valued, we use two channels to represent real- and imaginary part, respectively. The number of learned filters is K.

For the following experiments, we used a total number of $T = 25$ iterations of ALONE. For each iteration in the ALONE algorithm 1, the number of back-propagations for training the network $\mathbf{\Phi}_\theta$ to learn to patch-wise approximate the current image estimate was 400. The patch-size was chosen to be $32 \times 32 \times 4$. Network training was carried out by minimizing the loss function $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta)$ using the Adam optimizer [23] with a learning rate of 0.001. Further, before training, all input patches were normalized by subtracting the mean and dividing the patch by the standard deviation. For the reassembling of the patches, the normalization is reversed after having processed them with the network $\mathbf{\Phi}_\theta$. We set the number of learned filters to $K = 16$ and used the L_2-norm of the learned kernels as parameter regularization $\Omega(\theta)$ used in (4). When given $\mathbf{c}_k$, the system $\mathbf{H}\mathbf{x} = \mathbf{c}_k$ was solved by performing $n_{\text{iter}} = 4$ iterations of PCG.

C. Comparison to Other Iterative Methods

In this Section, we compare our proposed reconstruction method to the well known total variation-minimization algorithm [27] which has been successfully applied to $2D$ cine MRI [7] and to an iterative reconstruction algorithm based on learned dictionaries [9], [10], which we abbreviate by TV and DIC, respectively. Note that the methods in [9] and [10] further include a total variation penalty term in the formulation of the reconstruction problem which was reported to further increase the image quality of the reconstruction. However, in order to better compare the effect of the differently learned components

of the reconstruction algorithms, we neglect the TV-penalty term for the dictionary learning-based reconstruction. Figure 4 shows an example of results obtained by the three different methods.

1) Total Variation-Minimization: Our first method of comparison is the well known TV-minimization-based reconstruction, see for example [7] or [28] and [9], in the case the dictionary learning-based regularization term is neglected. The reconstruction problem is formulated as

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}_I\mathbf{x} - \mathbf{y}_I\|_2^2 + \frac{\lambda}{2}\|\mathbf{G}\mathbf{x}\|_1, \tag{10}$$

where $\mathbf{G}$ denotes the discretized version of the isotropic first order finite differences filter in all three dimensions. Problem (10) is solved by ADMM as in [9], by introducing an auxiliary variable $\mathbf{z}$ and alternatively solving for $\mathbf{z}$ and $\mathbf{x}$. For updating $\mathbf{z}$, an iterative shrinkage method is used, see [27]. Updating $\mathbf{x}$ corresponds to solving a problem which is linear in $\mathbf{x}$ and therefore to solving a system of linear equations, for which we used the pre-conditioned conjugate gradient method (PCG). We used a total number of $n_{\text{iter}} = 16$ iterations for ADMM, where $\mathbf{z}$ is updated using one iteration of the iterative shrinkage method and the linear system for the second subproblem for updating $\mathbf{x}$ is solved by $n_{\text{iter}} = 4$ iterations of PCG.

2) Dictionary Learning-based Regularization: The DIC method for comparison is given by the iterative reconstruction scheme using spatio-temporal learned dictionaries as regularizers presented in [9], [10]. We used the method by neglecting the TV-penalty term. This means the reconstruction problem is formulated as

$$\min_{\mathbf{x},\mathbf{D},\{\boldsymbol{\gamma}_j\}_j} \|\mathbf{A}_I\mathbf{x} - \mathbf{y}_I\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^{p}\|\mathbf{E}_j(\mathbf{x}) - \mathbf{D}\boldsymbol{\gamma}_j\|_2^2, \tag{11}$$

where $\mathbf{E}_j$ is again a patch-extraction operator, $\mathbf{D}$ is a dictionary and $\{\boldsymbol{\gamma}_j\}_j$ is a family of sparse codes. Problem (11) is also solved via ADMM by alternating the update with respect to $\mathbf{x}$ and the dictionary $\mathbf{D}$ as well as the sparse codes $\boldsymbol{\gamma}_j$. For learning the dictionary from the current estimate of the solution $\mathbf{x}$, we performed 10 iterations of the iterative thresholding an K residual means method[29], which is a faster alternative to K-SVD [30], while for obtaining the sparse codes, we used orthogonal matching pursuit [31]. As in [9] and [10], the dictionary was trained on patches of shape $4 \times 4 \times 4$. However, in contrast to the original works, we found a sparsity level $S = 16$ and a number of atoms of $K = d = 64 = 4 \cdot 4 \cdot 4$ to deliver more accurate results. This can most probably be related to the fact that, in contrast to [10] and [9], our k-space data is acquired along radial trajectories and the undersampling artefacts have an inherently different structure from the ones obtained by Cartesian sampling.

D. Results

Table I summarizes the results obtained by the two just introduced methods of comparison and our proposed approach.

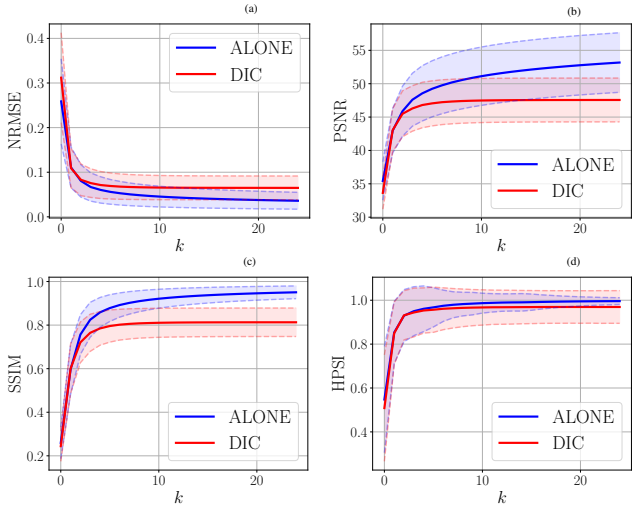

Figure 3. Convergence behaviour of the different learning-based methods with respect to the reported quantitative measures. The solid lines correspond to the mean value of the statistics calculated over the complete dataset. The dashed lines denote given by the average measured $\pm$ the corresponding standard deviation.

The Table was obtained by averaging the measures over all $N_z = 36$ slices of the all four patients which we reconstructed for all methods. The first column shows the measures corresponding to the NUFFT-reconstruction which is directly obtained from the measured k-space data. The second column shows the results obtained by the TV-minimization approach which increases image quality with respect to all reported measures. The DIC method and our approach further improve the image quality as can be seen by a further increase of SSIM, HPSI and PSNR and decrease of NRMSE. However, our proposed reconstruction considerably outperforms DIC by ≈ 5.3 dB in terms of PSNR, $\approx 7\%$ in terms of SSIM and 0.024 in terms of NRMSE. The increase of HPSI on the other hand, is relatively small.

Since our proposed method and the DIC method approach are similar in the sense that the regularization is adaptively learned from the current image estimate during the iterative reconstruction, we investigated the convergence behaviour of the two methods. Figure 3 shows different curves for the different reported quantitative measures during the reconstruction. The solid lines correspond to the mean value of the measure averaged over the complete dataset of $N_z = 36$ slices. Further, the dashed lines show the curves given by the mean $\pm$ the standard deviation of the considered measure. The measures were calculated after having solved the system $\mathbf{H}\mathbf{x} = \mathbf{c}_k$, where $\mathbf{c}_k$ is given as in (8) for our method. For the DIC method, $\mathbf{c}_k$ is given as in (8) but with $\mathbf{z}_j$ given as the sparse approximation of all patches of the current image estimate, i.e. $\mathbf{z}_j = \mathbf{D}\gamma_j$ for all j. Note that, while the DIC methd reaches a point of stagnation in terms of NRMSE decrease (which naturally corresponds the measure which is minimized during the iterative reconstruction) and PSNR between the fifth and tenth iteration, our proposed method ALONE seems to still have the potential to further improve image image quality, as neither NRMSE nor PSNR or SSIM have reached a point of

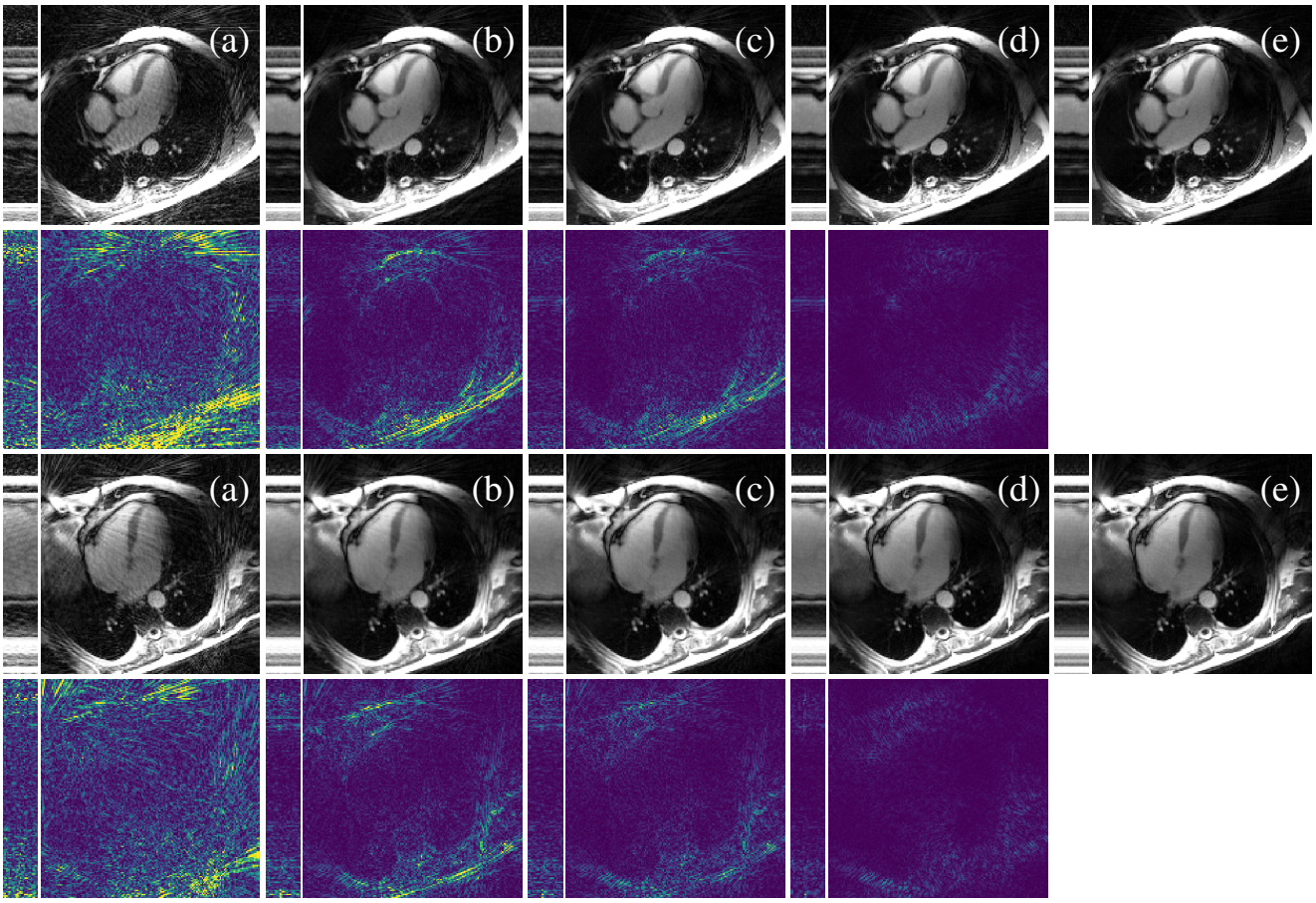

Figure 4. Results obtained by different methods based on iterative reconstruction. NUFFT reconstruction from $N_\varphi = 1130$ radial spokes (a), TV minimization-based regularization (b), DIC-based regularization (c), proposed regularization based on shallow CNNs (d), kt-SENSE reconstruction from $N_\varphi = 3400$ radial spokes (e).

saturation. Further, note how for all measures except for PSNR the standard deviation of the measure becomes smaller during the reconstruction which indicates an improved stability of the algorithm compared to the DIC method.

Table I
COMPARISON ALONE WITH DIFFERENT ITERATIVE METHODS USING DIFFERENT REGULARIZATIONS.

	NUFFT	**TV**	**DIC**	**ALONE**
PSNR	35.496	40.412	42.858	48.122
SSIM	0.626	0.838	0.895	0.962
HPSI	0.955	0.981	0.991	0.998
NRMSE	0.137	0.077	0.058	0.033

IV. DISCUSSION

In this work, we have presented a simple yet powerful method named ALONE (Adaptive Learning of NEtworks) for regularization for 2D undersampled radial cine MRI. The method is based on the adaptive regularization of the solution given in the form of a shallow CNN which is trained in an unsupervised manner during the reconstruction.

We have compared ALONE to a well-known total variation-minimization approach (TV) as well as to another learning-based method which employs an adaptive regularization based on dictionary learning (DIC). Our method outperforms the TV- and the DIC method with respect to all reported measures. Further, we investigated the effect of the adaptive regularization for both learning-based methods during the reconstruction. ALONE shows an improved and more stable convergence behaviour during the reconstruction which is visible in terms of a smaller standard deviation of NRMSE, SSIM and HPSI during the reconstruction.

Further, our proposed approach ALONE has one significant advantage over the dictionary learning-based method DIC, which is the acceleration of the regularization step during the reconstruction. Note from (11), that the reconstruction problem is formulated as joint minimization problem over the variables $\mathbf{x}$, $\mathbf{D}$ and $\{\boldsymbol{\gamma}_j\}$. In contrast, from (1) we see that our formulation only requires the update of two variables. While for DIC, training the dictionary $\mathbf{D}$ is achieved in a relatively short time, the computational bottleneck of the approach is finding the sparse codes $\boldsymbol{\gamma}_j$ of the patches $\mathbf{E}_j(\mathbf{x})$ with respect to the dictionary $\mathbf{D}$. This is because obtaining $\boldsymbol{\gamma}_j$ involves solving an optimization problem for all j, namely the sparse coding problem. On the other hand, the sparse-

approximation counterpart in our reconstruction scheme is given by calculating $\mathbf{z}_j = \mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)$, i.e. by performing a forward pass of the patches through the (shallow) CNN. Table II shows a direct comparison of the corresponding counterparts for the DIC method and our proposed reconstruction scheme ALONE. The time in the Table refers to the average time needed for obtaining the regularized image, i.e. for the patch-wise sparse approximation using OMP for the DIC method, and for obtaining $\mathbf{c}_k$ for our method ALONE by performing a forward pass of all patches. Therefore, the total cost of the regularization can be estimated by multiplying the average time by the number of iterations T one sets before the reconstruction.

Table II
COMPARISON OF THE DIFFERENT COMPONENTS FOR THE REGULARIZATION WITH DICTIONARY LEARNING AND OUR PROPOSED METHOD ALONE.

	DIC	**ALONE**
Patches size	$4 \times 4 \times 4$	$32 \times 32 \times 4$
Strides	$2 \times 2 \times 2$	$16 \times 16 \times 2$
Number of patches	353 934	5 054
Training	ITKrM	Back-propagation
Time	$\approx$ 10 s	$\approx$ 6 s
Patches approximation	OMP	Forward pass
Time	$\approx$ 7 m	$\approx$ 0.3 s
Data type of patches	$\mathbb{R}$	$\mathbb{C}$

In [9], it was reported that for the DIC method, using real-valued dictionaries for the sparse approximation of the complex-valued images outperformed the usage of complex-valued dictionaries. Note that this adds another (non-negligible) factor of two to the most computational demanding component of the reconstruction, i.e. the sparse coding of all patches. In our proposed reconstruction scheme ALONE, in contrast, there is no noticeable difference between using a real-valued and a complex-valued CNN in terms of speed, as the additional increase of complexity is negligible. For ALONE, training the CNN on complex-valued patches represented by two input-channels yielded more accurate reconstruction. Since iterative reconstruction is time consuming, we only reconstructed the image data for only one of the four patients with $N_z = 12$ slices by employing a CNN-based regularization which is learned from the real-valued patches. Similarly to [9], the complex-valued patches were then obtained by performing a forward pass of the real and the imaginary part of the patches using the same CNN. Figure 5 shows an example of results obtained with ALONE using a real-valued and complex-valued CNN, where we see that the complex-valued CNN improved the results. Further, from the point-wise error images and the yellow arrows, we see that DIC tends to slightly smooth image details, while ALONE well preserves edges. Table III shows the obtained results for one of the patients for the real- and the complex version of our proposed ALONE reconstruction as well as for the DIC method.

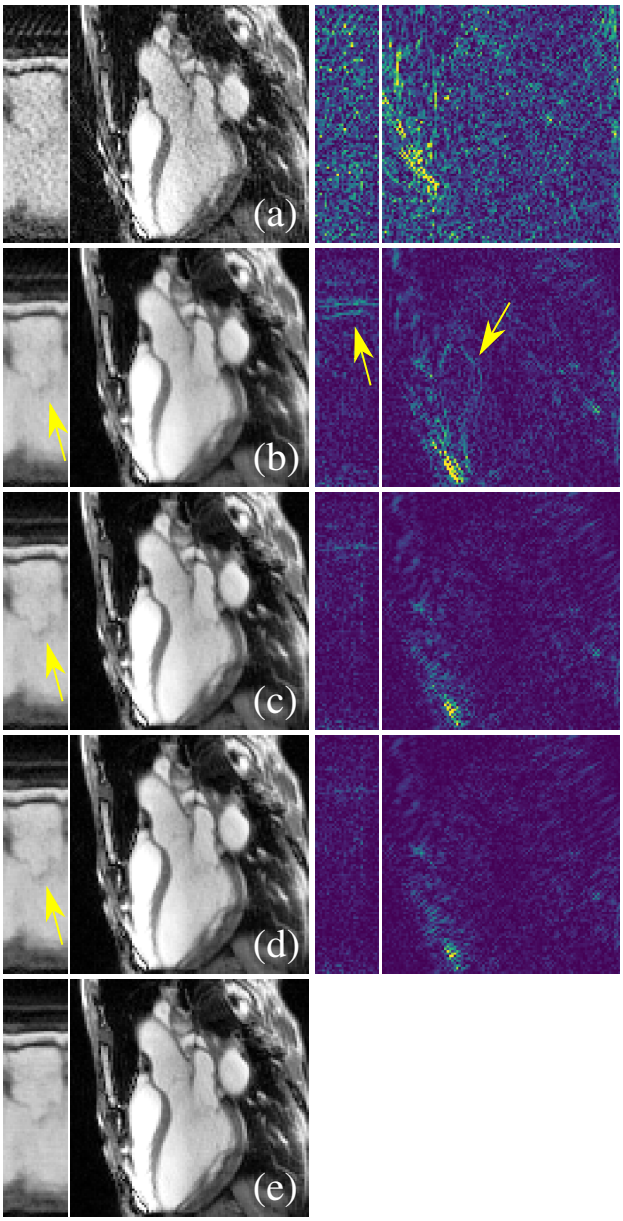

Figure 5. Results obtained by the NUFFT reconstruction from $N_\varphi = 1130$ radial spokes (a), with the DIC method (b), with ALONE by using a complex-valued CNN $\mathbf{\Phi}_\theta$ (c), with ALONE by using a real-valued CNN $\mathbf{\Phi}_\theta$ (d) and the reference kt-SENSE reconstruction obtained from $N_\varphi = 3400$ radial spokes.

Table III
COMPARISON OF ALONE USING A REAL-VALUED AND COMPLEX-VALUED CNN $\mathbf{\Phi}_\theta$ TO DIC USING REAL VALUED DICTIONARIES.

	DIC ($\mathbb{R}$)	**ALONE** ($\mathbb{R}$)	**ALONE** ($\mathbb{C}$)
PSNR	40.623	43.669	45.505
SSIM	0.868	0.926	0.950
HPSI	0.994	0.998	0.999
NRMSE	0.059	0.043	0.035

Another advantage of using ALONE is that we can use relatively large patch-sizes and strides without observing artefacts which could be attributed to the overlapping of patches. The reason for that most probably lies in the fact that the network $\mathbf{\Phi}_\theta$ is a CNN. Since CNNs are well-known to be translation-equivariant, using larger strides for the regularization of the solution does not lead to block-artefacts. In contrast, since the approximation using a learned dictionary can be identified with a single fully-connected layer taking the sparse code $\boldsymbol{\gamma}_j$ as an input, the operation is not translation-equivariant in general and therefore, relatively small strides need to be used in order to avoid patchy artefacts. Compared to the DIC method, in our proposed method ALONE, the number of patches needed to be processed is lower, larger strides can be used and the patch-wise approximation is faster. In sum, this results in an acceleration of the regularization step by several orders of magnitude. Therefore, for ALONE, the overall cost of the reconstruction is dominated by the application of the forward and adjoint operators as in any other reconstruction algorithm.

Even if the regularization of the solution in each iteration is highly accelerated, the main limitation of the method clearly remains the relatively high number of iterations needed to perform the reconstruction. Further, the strength of the regularization λ has to be chosen a priori which might be problem-dependent. However, note that the usage of cascaded networks, which can be thought of unrolled iterative schemes is prohibitive for large-scale problems as the one considered in this work. Opposed to other works using cascaded networks for the MR image reconstruction task, see e.g. [18], [17], [19], our forward operator $\mathbf{A}_I$ is given by a radial Fourier encoding operator with $n_c = 12$ coils and does not allow the construction of iterative neural networks due to its computational complexity. For a more detailed discussion about the issue and a possible way to overcome the problem, we refer to [14].

V. Conclusion

In this work we have presented a new reconstruction algorithm named for accelerated 2D radial cine MRI. The reconstruction algorithm involves a patch-wise regularization which is adaptively learned during the reconstruction in an unsupervised manner. Therefore, the method does not require having access to large training datasets with ground truth data.

We have compared our reconstruction method to a total variation-minimization method and to a dictionary learning-based method using adaptively trained dictionaries. Our method outperformed both methods with respect to all chosen reported measures. Further, compared to the dictionary learning-based method, it accelerates the reconstruction by orders of magnitude since it highly reduces the regularization step needed during the reconstruction.

While in this work we have applied our reconstruction method ALONE to 2D cardiac radial cine MRI, the method's formulation is held general and therefore we expect ALONE to be applicable to other imaging modalities as well.

Acknowledgements

A. Kofler and M. Dewey acknowledge the support of the German Research Foundation (DFG), project number GRK 2260, BIOQIC. The work of M. Haltmeier has been supported by the Austrian Science Fund (FWF), project P 30747-N32.

Appendix
Mathematical Analysis of ALONE

In this appendix we present some theoretical results for Algorithm 1 (ALONE). For that purpose, we assume absence of noise in the measurements and note that ALONE is of the fixed point form

$$\theta_k \in \arg\min_\theta \frac{\lambda}{2} \left\| \mathbf{E}(\mathbf{x}_k) - \mathbf{\Phi}_\theta \mathbf{E}(\mathbf{x}_k) \right\|_2^2 + \Omega(\theta)\,,$$
$$\mathbf{x}_{k+1} \in (\mathbf{A}_I^\mathbf{H}\mathbf{A}_I + \lambda \mathbf{E}^\mathbf{T}\mathbf{E})^{-1}(\mathbf{A}_I^\mathbf{H}\mathbf{y}_I + \lambda \mathbf{E}^\mathbf{T}\mathbf{\Phi}_{\theta_k}\mathbf{E}\mathbf{x}_k)\,,$$

with initialization $\mathbf{x}_0 = \mathbf{A}_I^\mathbf{H}\mathbf{y}_I$.

A. Characterization of fixed points

We first define the natural underlying prior information for solutions of (2) induced by ALONE.

Definiton A.1 (θ^*-adapted solution): For any $\theta^* \in \mathbb{R}^q$, we call $\mathbf{x}^* \in \mathbb{C}^N$ a θ^*-adapted solution of (2), if

$$\mathbf{A}_I \mathbf{x}^* = \mathbf{y}_I \tag{12}$$
$$\mathbf{E}\mathbf{x}^* = \mathbf{\Phi}_{\theta^*}\mathbf{E}\mathbf{x}^* \tag{13}$$

We will show that θ^*-adapted solutions are fixed points of the ALONE algorithm as well as partial minimizers of $\mathcal{R}_{\lambda,\mathbf{y}_I}$ defined next.

Definiton A.2 (Fixed points): The pair $(\theta^*, \mathbf{x}^*) \in \mathbb{C}^N \times \mathbb{R}^q$ is called fixed point of ALONE if

$$\theta^* \in \arg\min_\theta \frac{\lambda}{2} \left\| \mathbf{E}\mathbf{x}^* - \mathbf{\Phi}_\theta(\mathbf{E}\mathbf{x}^*) \right\|_\mathbf{E}^2 + \Omega(\theta^*) \tag{14}$$
$$\mathbf{x}^* \in (\mathbf{A}_I^\mathbf{H}\mathbf{A}_I + \lambda \mathbf{E}^\mathbf{T}\mathbf{E})^{-1}(\mathbf{A}_I^\mathbf{H}\mathbf{y}_I + \lambda \mathbf{E}^\mathbf{T}\mathbf{\Phi}_{\theta^*}\mathbf{E}\mathbf{x}^*)\,. \tag{15}$$

Definiton A.3 (Partial minimizers): The pair $(\mathbf{x}^*, \theta^*) \in \mathbb{C}^N \times \mathbb{R}^q$ is called a partial minimizer of $\mathcal{R}_{\lambda,\mathbf{y}_I}$ if

$$\forall \mathbf{x} \in \mathbb{C}^N\colon \quad \mathcal{R}_{\lambda,\mathbf{y}_I}(\mathbf{x}^*, \theta^*) \leq \mathcal{R}_{\lambda,\mathbf{y}_I}(\mathbf{x}, \theta^*) \tag{16}$$
$$\forall \theta \in \mathbb{R}^q\colon \quad \mathcal{R}_{\lambda,\mathbf{y}_I}(\mathbf{x}^*, \theta^*) \leq \mathcal{R}_{\lambda,\mathbf{y}_I}(\mathbf{x}^*, \theta) \tag{17}$$

We have the following result relating θ^*-adapted solutions of inverse problems of the form given in (2) to fixed points of ALONE and partial minimizers of $\mathcal{R}_{\lambda,\mathbf{y}_I}$.

Theorem A.4 (Fixed points and partial minimizers for θ^*-adapted solutions): Let $\theta^* \in \mathbb{R}^q$ satisfy (14) and $\mathbf{x}^* \in \mathbb{C}^N$ be a θ^*-adapted solution of (2). Then the following hold:

(a) $(\theta^*, \mathbf{x}^*)$ is a fixed point of ALONE.

(b) $(\theta^*, \mathbf{x}^*)$ is a partial minimizer of $\mathcal{R}_{\lambda,\mathbf{y}_I}$.

Proof: (a) Equation (14) is satisfied by assumption. Further, (15) is satisfied if and only if the optimality condition

$$0 = \lambda(\mathbf{E}^\mathsf{T}\mathbf{E}\mathbf{x}^* - \mathbf{E}^\mathsf{T}\mathbf{\Phi}_{\theta^*}\mathbf{E}\mathbf{x}^*) + \mathbf{A}_I^\mathsf{H}(\mathbf{A}_I\mathbf{x}^* - \mathbf{y}_I)$$

holds. According to (12), (13) this is however the case. Hence $(\theta^*, \mathbf{x}^*)$ is a fixed point of ALONE.

(b) According to (12), (13) we have $\mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x}^*, \theta^*) = \Omega(\theta^*)$. Consequently, for any $\mathbf{x} \in \mathbb{C}^N$ we have

$$\begin{aligned}\mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x}, \theta^*) &= \frac{\lambda}{2}\left\|\mathbf{E}\mathbf{x} - \mathbf{\Phi}_{\theta^*}\mathbf{E}\mathbf{x}\right\|_2^2 \\ &+ \frac{1}{2}\left\|\mathbf{A}_I\mathbf{x} - \mathbf{y}_I\right\|_2^2 + \Omega(\theta^*) \geq \Omega(\theta^*) = \mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x}^*, \theta^*)\,.\end{aligned}$$

This shows (16). Using (14), one verifies

$$\begin{aligned}\mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x}^*, \theta) &= \frac{\lambda}{2}\left\|\mathbf{E}\mathbf{x}^* - \mathbf{\Phi}_{\theta}\mathbf{E}\mathbf{x}^*\right\|_2^2 + \Omega(\theta) \\ &\quad + \frac{1}{2}\left\|\mathbf{A}_I\mathbf{x}^* - \mathbf{y}_I\right\|_2^2 \geq \mathcal{R}_{\mathbf{y}_I,\lambda}(\mathbf{x}^*, \theta^*)\,.\end{aligned}$$

Hence (17) is satisfied for any $\theta \in \mathbb{R}^q$ and $(\theta^*, \mathbf{x}^*)$ is a partial minimizer of $\mathcal{R}_{\lambda,\mathbf{y}_I}$. ■

B. Existence and stability

Next we show that the ALONE algorithm 1 is well-defined (iterates exist) and stable with respect to data perturbation. For that purpose we assume that the following reasonable conditions hold.

Assumption A.5 (Existence and stability):

(A1) $\mathbf{A}_I \colon \mathbb{C}^N \to \mathbb{C}^m$ is a linear forward operator.

(A2) $\mathbb{R}^q \times \mathbb{C}^N \to \mathbb{C}^N \colon (\theta, \mathbf{x}) \mapsto \mathbf{\Phi}_\theta(\mathbf{x})$ is continuous.

(A3) $\forall j \in \{1, \dots, p\} \colon\ \mathbf{E}_j \colon \mathbb{C}^N \to \mathbb{C}^d$ is linear.

(A4) $\Omega \colon \mathbb{R}^q \to [0, \infty)$ is continuous and coercive.

All above assumptions are naturally fulfilled in our context. (A2) is satisfied for typical network architectures, in particular for CNNs. (A3) is satisfied for the patch extraction operator and, finally (A4) is satisfied for typically used regularizers such as the Frobenius-norm or weighted ℓ^q-norms. Recall that Ω is called coercive if $\Omega(\theta_k) \to \infty$ if $(\theta_k)_{k\in\mathbb{N}}$ is a sequence with $\|\theta_k\|_2 \to \infty$.

Theorem A.6 (Existence): Algorithm 1 defines a sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ of iterates and a sequence of parameters $(\theta_k)_{k\in\mathbb{N}}$.

Proof: In order to show that ALONE defines sequences $(\mathbf{x}_k)_{k\in\mathbb{N}}$, $(\theta_k)_{k\in\mathbb{N}}$, it is sufficient to show that the functionals $\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}$ and $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}$ both have at least one minimizer. Because of (A1), (A2), the functional $\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}$ is quadratic and nonnegative and therefore has a minimizer. According to (A2) the mapping $\theta \mapsto \mathbf{\Phi}_\theta(\mathbf{E}(\mathbf{x}))$ is continuous. Together with (A4) this implies that $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta) = \|\mathbf{E}\mathbf{x} - \mathbf{\Phi}_\theta(\mathbf{E}(\mathbf{x}))\|_2^2 + \Omega(\theta)$ is continuous and coercive. Standard arguments therefore imply the existence of minimizers. To see this, let $(\theta_k)_{k\in\mathbb{N}}$ be a sequence with $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta_k) \to \inf_\theta \mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}(\theta)$. The coercivity of $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}$ implies that this sequence is bounded and therefore has at least one convergent subsequence. The continuity implies that the limit of the subsequence is a minimizer of $\mathcal{L}_{\mathbf{x},\mathbf{y}_I}$. ■

Note that that the minimizer of $\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}$ is unique if the matrix $\mathbf{H}$ has full rank. For example, this is the case of $\mathbf{E}^\mathsf{T}\mathbf{E}$ has full rank. For the patch extraction operator, the full rank condition is satisfied, provided the patches cover the whole image domain. However, the minimizer of $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}$ might be non-unique due to the non-convexity of $\theta \mapsto \|\mathbf{E}_j(\mathbf{x}) - \mathbf{\Phi}_\theta\big(\mathbf{E}_j(\mathbf{x})\big)\|_2^2$. Therefore ALONE depends on the particular choice of the minimizers of $\mathcal{L}_{\mathbf{x},\mathbf{y}_I,\lambda}$ and, in general, does not define a unique sequence.

Definiton A.7 (Sets of ALONE iterates): For any $\mathbf{y}_I \in \mathbb{C}^m$ and any iteration index $k \in \mathbb{N}$ we denote by $\mathbf{B}_k(\mathbf{y}_I)$ the set of all iterates $(\mathbf{x}_k, \theta_k) \in \mathbb{C}^N \times \mathbb{R}^q$ that are generated by the ALONE reconstruction Algorithm 1, by choosing arbitrary minimizers of $\mathcal{L}_{\mathbf{x}_\ell,\mathbf{y}_I,\lambda}$, $\mathcal{L}_{\theta_\ell,\mathbf{y}_I,\lambda}$ for $\ell \in \{0, \dots, k-1\}$.

We next show that the sets of ALONE iterates $\mathbf{B}_k(\mathbf{y}_I)$ depend stably on the inputs $\mathbf{y}_I$. For that purpose, we write $\mathbf{B} \colon \mathbb{C}^m \rightrightarrows \mathbb{C}^N \times \mathbb{R}^q$ for a multivalued mappings where $\mathbf{B}(\mathbf{y}_I) \subset \mathbb{C}^N \times \mathbb{R}^q$ and use the following notion of stability.

Definiton A.8 (Stability of multivalued mappings): A multivalued mapping $\mathbf{B} \colon \mathbb{C}^m \rightrightarrows \mathbb{C}^N \times \mathbb{R}^q$ is called stable if for any $\mathbf{y}_I \in \mathbb{C}^m$ and any sequence $(\mathbf{y}_I^\ell)_{\ell\in\mathbb{N}}$ converging to $\mathbf{y}_I$ the following statements hold true:

(i) $(\mathbf{B}(\mathbf{y}_I^\ell))_{\ell\in\mathbb{N}}$ has at least one accumulation point.

(ii) All accumulation points of $(\mathbf{B}(\mathbf{y}_I^\ell))_{\ell\in\mathbb{N}}$ are in $\mathbf{B}(\mathbf{y}_I)$.

Theorem A.9 (Stability of ALONE iterates): For any iteration index $k \in \mathbb{N}$, the set of ALONE iterates $\mathbf{B}_k$ is stable.

Proof: An inductive argument shows that it is sufficient verify that $\mathbf{x} \mapsto \arg\min_\theta \mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x})$ is stable. For that purpose, let $(\mathbf{y}_I^\ell)_{\ell\in\mathbb{N}}$ be a sequence converging to $\mathbf{y}_I \in \mathbb{C}^m$ and let $\mathbf{x}^* \in \arg\min \mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x})$, $\mathbf{x}^\ell \in \arg\min \mathcal{L}_{\theta,\mathbf{y}_I^\ell,\lambda}(\mathbf{x})$. Then $\mathcal{L}_{\theta,\mathbf{y}_I^\ell,\lambda}(\mathbf{x}^\ell) \leq \mathcal{L}_{\theta,\mathbf{y}_I^\ell,\lambda}(\mathbf{x}^*)$ and therefore

$$\begin{aligned}&\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x}^\ell) \\ &\quad= \frac{\lambda}{2}\left\|\mathbf{E}\mathbf{x}^\ell - \mathbf{Z}\right\|_2^2 + \frac{1}{2}\left\|\mathbf{A}_I\mathbf{x} - \mathbf{y}_I\right\|_2^2 \\ &\quad\leq \frac{\lambda}{2}\left\|\mathbf{E}\mathbf{x}^\ell - \mathbf{Z}\right\|_2^2 + \left\|\mathbf{A}_I\mathbf{x}^\ell - \mathbf{y}_I^\ell\right\|_2^2 + \left\|\mathbf{y}_I - \mathbf{y}_I^\ell\right\|_2^2 \\ &\quad\leq 2\mathcal{L}_{\theta,\mathbf{y}_I^\ell,\lambda}(\mathbf{x}^\ell) + \left\|\mathbf{y}_I - \mathbf{y}_I^\ell\right\|_2^2 \\ &\quad\leq 2\mathcal{L}_{\theta,\mathbf{y}_I^\ell,\lambda}(\mathbf{x}^*) + \left\|\mathbf{y}_I - \mathbf{y}_I^\ell\right\|_2^2 \\ &\quad\leq 4\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x}^*) + 3\left\|\mathbf{y}_I - \mathbf{y}_I^\ell\right\|_2^2\,.\end{aligned}$$

Because $\left\|\mathbf{y}_I - \mathbf{y}_I^\ell\right\|_2 \to 0$, this and the coercivity of $\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}$ show that $(\mathbf{x}^\ell)_{\ell\in\mathbb{N}}$ is bounded. In particular, $(\mathbf{x}^\ell)_{\ell\in\mathbb{N}}$ has at least one accumulation point. Let $(\mathbf{x}^{\tau(\ell)})_{\ell\in\mathbb{N}}$ be a subsequence converging to some $\hat{\mathbf{x}} \in \mathbb{C}^N$. The continuity of the norm implies

$$\begin{aligned}&\mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\hat{\mathbf{x}}) = \lim_{\ell\to\infty} \mathcal{L}_{\theta,\mathbf{y}_I^{\tau(\ell)},\lambda}(\mathbf{x}^{\tau(\ell)}) \\ &\leq \liminf_{\ell\to\infty} \mathcal{L}_{\theta,\mathbf{y}_I^{\tau(\ell)},\lambda}(\mathbf{x}^*) = \mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x}^*) = \inf_{\mathbf{x}} \mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x})\,.\end{aligned}$$

Consequently, $\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}} \mathcal{L}_{\theta,\mathbf{y}_I,\lambda}(\mathbf{x})$. ■

Clearly a main theoretical questions is to show, under suitable assumptions, convergence of ALONE to fixed points defined by (14), (15). This turned out to be a very challenging question that we aim to investigate in future work.

REFERENCES

[1] M. Weiger, K. P. Pruessmann, and P. Boesiger, "Cardiac real-time imaging using sense," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 43, no. 2, pp. 177–184, 2000.

[2] F.-H. Lin, K. K. Kwong, J. W. Belliveau, and L. L. Wald, "Parallel imaging reconstruction using automatic regularization," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 51, no. 3, pp. 559–567, 2004.

[3] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[4] D. L. Donoho *et al.*, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] N. Seiberlich, G. Lee, P. Ehses, J. L. Duerk, R. Gilkeson, and M. Griswold, "Improved temporal resolution in cardiac imaging using through-time spiral grappa," *Magnetic Resonance in Medicine*, vol. 66, no. 6, pp. 1682–1688, 2011.

[6] S. Winkelmann, T. Schaeffter, T. Koehler, H. Eggers, and O. Doessel, "An optimal radial profile order based on the golden ratio for time-resolved mri," *IEEE Transactions on Medical Imaging*, vol. 26, no. 1, pp. 68–76, 2006.

[7] K. T. Block, M. Uecker, and J. Frahm, "Undersampled radial mri with multiple coils. iterative image reconstruction using a total variation constraint," *Magnetic Resonance in Medicine*, vol. 57, no. 6, pp. 1086–1098, 2007.

[8] S. Ravishankar and Y. Bresler, "Mr image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Transactions on Medical Imaging*, vol. 30, no. 5, pp. 1028–1041, 2010.

[9] J. Caballero, A. N. Price, D. Rueckert, and J. V. Hajnal, "Dictionary learning and time sparsity for dynamic mr data reconstruction," *IEEE Transactions on Medical Imaging*, vol. 33, no. 4, pp. 979–994, 2014.

[10] Y. Wang and L. Ying, "Compressed sensing dynamic cardiac cine mri using learned spatiotemporal dictionary," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 4, pp. 1109–1120, 2014.

[11] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, "Deep Convolutional Neural Network for Inverse Problems in Imaging," *IEEE Transactions on Image Processing*, 2017.

[12] Y. Han and J. C. Ye, "Framing u-net via deep convolutional framelets: Application to sparse-view ct," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1418–1429, 2018.

[13] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J. A. Steeden, "Real-time cardiovascular mr with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease," *Magnetic Resonance in Medicine*, vol. 81, no. 2, pp. 1143–1156, 2019.

[14] A. Kofler, M. Dewey, T. Schaeffter, C. Wald, and C. Kolbitsch, "Spatio-temporal deep learning-based undersampling artefact reduction for 2d radial cine mri with limited training data," *IEEE Transactions on Medical Imaging*, no. DOI: 10.1109/TMI.2019.2930318, 2019.

[15] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," *Inverse Problems*, vol. 33, no. 12, p. 124007, 2017.

[16] ——, "Learned primal-dual reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1322–1332, 2018.

[17] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *Magnetic Resonance in Medicine*, vol. 79, no. 6, pp. 3055–3071, 2018.

[18] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 37, no. 2, pp. 491–503, 2018.

[19] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueckert, "Convolutional recurrent neural networks for dynamic mr image reconstruction," *IEEE Transactions on Medical Imaging*, vol. 38, no. 1, pp. 280–290, 2018.

[20] L. Feng, L. Axel, H. Chandarana, K. T. Block, D. K. Sodickson, and R. Otazo, "XD-GRASP : Golden-Angle Radial MRI with Reconstruction of Extra Motion-State Dimensions Using Compressed Sensing," *Magn. Reson. Med.*, vol. 00, no. October 2014, pp. 1–14, 2015.

[21] L. Feng, M. B. Srichai, R. P. Lim, A. Harrison, W. King, G. Adluru, E. V. R. Dibella, D. K. Sodickson, R. Otazo, and D. Kim, "Highly accelerated real-time cardiac cine MRI using k-t SPARSE-SENSE." *Magn. Reson. Imag.*, aug 2012. [Online]. Available: http://dx.doi.org/10.1002/mrm.24440

[22] J.-M. Lin, "Python non-uniform fast fourier transform (pynufft): An accelerated non-cartesian mri package on a heterogeneous platform (cpu/gpu)," *Journal of Imaging*, vol. 4, no. 3, p. 51, 2018.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] M. R. Hestenes, E. Stiefel *et al.*, "Methods of conjugate gradients for solving linear systems," *Journal of research of the National Bureau of Standards*, vol. 49, no. 6, pp. 409–436, 1952.

[25] J. Tsao, P. Boesiger, and K. P. Pruessmann, "k-t blast and k-t sense: dynamic mri with high frame rate exploiting spatiotemporal correlations," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 50, no. 5, pp. 1031–1042, 2003.

[26] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, "A haar wavelet-based perceptual similarity index for image quality assessment," *Signal Processing: Image Communication*, vol. 61, pp. 33–43, 2018.

[27] A. Chambolle, "Total variation minimization and a class of binary mrf models," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2005, pp. 136–152.

[28] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[29] K. Schnass, "Convergence radius and sample complexity of itkm algorithms for dictionary learning," *Applied and Computational Harmonic Analysis*, vol. 45, no. 1, pp. 22–58, 2018.

[30] M. Aharon, M. Elad, A. Bruckstein *et al.*, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, p. 4311, 2006.

[31] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.