Learning-Based Compressive MRI

Baran Gözü¹, Rabeeh Karimi Mahabadi¹, Yen-Huan Li¹, Efe Ilıcak², Tolga Çukur²²³, Jonathan Scarlett⁴, and Volkan Cevher¹

¹Laboratory for Information and Inference Systems (LIONS), EPFL, Switzerland
²National Magnetic Resonance Research Center (UMRAM), Bilkent University, Ankara, Turkey
³Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey
⁴Department of Computer Science & Department of Mathematics, National University of Singapore, Singapore

In the area of magnetic resonance imaging (MRI), an extensive range of non-linear reconstruction algorithms have been proposed that can be used with general Fourier subsampling patterns. However, the design of these subsampling patterns has typically been considered in isolation from the reconstruction rule and the anatomy under consideration. In this paper, we propose a learning-based framework for optimizing MRI subsampling patterns for a specific reconstruction rule and anatomy, considering both the noiseless and noisy settings. Our learning algorithm has access to a representative set of training signals, and searches for a sampling pattern that performs well on average for the signals in this set. We present a novel parameter-free greedy mask selection method, and show it to be effective for a variety of reconstruction rules and performance metrics. Moreover we also support our numerical findings by providing a rigorous justification of our framework via statistical learning theory.

Index Terms—Magnetic resonance imaging, compressive sensing, learning-based subsampling, greedy algorithms

I. INTRODUCTION

Magnetic resonance imaging (MRI) serves as a crucial diagnostic modality for scanning soft tissue in body parts such as the brain, knee, and spinal cord. While early MRI technology could require over an hour of scan time to produce diagnostic-quality images, subsequent advances have led to drastic reductions in the scan time without sacrificing the imaging quality.

The application of MRI has served as a key motivation for compressive sensing (CS), a modern data acquisition technique for sparse signals. The theory and practice of CS for MRI have generally taken very different paths, with the former focusing on sparsity and uniform random sampling of the Fourier space, but the latter dictating the use of variable-density subsampling. Common to both viewpoints, however, is the element of non-linear decoding via optimization formulations.

In this paper, we propose a learning-based framework for compressive MRI that is both theoretically grounded and practical. The premise is to use training signals to optimize the subsampling specifically for the setup at hand.

In more detail, we propose a novel greedy algorithm for mask optimization that can be applied to arbitrary reconstruction rules and performance measures. This mask selection algorithm is parameter-free, excluding unavoidable parameters of the reconstruction methods themselves. We use statistical learning theory to justify the core idea of optimizing the empirical performance on training data for the sampling design problem. In addition, we provide numerical evidence that our framework can find good sampling patterns for different performance metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index [1], and for a broad range of decoders, from basis pursuit and total variation to neural networks and BM3D. Since our framework can be applied to arbitrary decoders, we also anticipate that it can benefit future decoding rules.

Organization of the paper: In Section II, we introduce the compressive MRI problem and outline the most relevant existing works, as well as summarizing our contributions. In Section III, we introduce our learning-based framework, along with its theoretical justification. In Section IV, we demonstrate the effectiveness of our approach on a variety of data sets, including comparisons to existing approaches. Conclusions are drawn in Section V.

II. BACKGROUND

A. Signal acquisition and reconstruction

In the compressive sensing (CS) problem [2], one seeks to recover a sparse vector via a small number of linear measurements. In the special case of compressive MRI, these measurements take the specific form of subsampled Fourier measurements, described as follows:

\[
b = P_\Omega \Psi x + w, \tag{1}\]

where \( \Psi \in \mathbb{C}^{p \times p} \) is the Fourier transform operator applied to the vectorized image, \( P_\Omega : \mathbb{C}^p \to \mathbb{C}^n \) is a subsampling operator that selects the rows of \( \Psi \) indexed by the set \( \Omega \), with \( |\Omega| = n \), and \( w \in \mathbb{C}^n \) is possible additive noise. We refer to \( \Omega \) as the sampling pattern or mask.

Given the measurements \( b \) (along with knowledge of \( \Omega \)), a reconstruction algorithm (also referred to as the decoder) forms an estimate \( \hat{x} \) of \( x \). This algorithm is treated as a general function, and is written as follows:

\[
\hat{x} = g(\Omega, b). \tag{2}\]

A wide variety of decoding techniques have been proposed for compressive MRI; here we present a few of the most widely-used and best-performing techniques, which we will pursue in the numerical experiments in Section IV.

¹The original image may be 2D or 3D, but we express it in its vectorized form for convenience.
In the general CS problem, decoders based on convex optimization have received considerable attention, both due to their theoretical guarantees and practical performance. In the noiseless setting (i.e., w = 0), a particularly notable choice is basis pursuit (BP) [2]:

$$\hat{x} = \arg\min_{z \in \mathbb{R}^n} \|\Phi z\|_1$$

(3)

where $\Phi$ is a sparsifying operator such as the wavelet or shearlet transform. A similar type of convex optimization formulation that avoids the need for the sparsifying operator is total variation (TV) minimization:

$$\hat{x} = \arg\min_{z \in \mathbb{R}^n} \|z\|_{TV},$$

(4)

where $\|z\|_{TV}$ is the total variation norm.

For the specific application of MRI, heuristic reconstruction algorithms have recently arisen that can outperform methods such as BP and TV, despite their lack of theoretical guarantees. A state-of-the-art reconstruction algorithm was recently proposed in [3] based on the block matching and 3D filtering (BM3D) denoising technique [4] that applies principal component analysis (PCA) to patches of the image. At a high level, the algorithm of [3] alternates between denoising using BM3D, and reconstruction using regularized least squares formulations. We refer the reader to [3] for further details, and to Section IV for our numerical results.

Following their enormous success in machine learning applications, deep neural networks have also been proposed for MRI reconstruction. We consider the approach of [5], which uses a cascade of convolutional neural networks (CNNs) interleaved with data consistency (DC) units. The CNNs serve to perform de-aliasing, and the DC units serve to enforce data consistency in the reconstruction. The deep network is trained by inputting the subsampled signals and treating the full training signal as the desired reconstruction. We refer the reader to [5] for further details, and to Section IV for our numerical results.

Among the extensive existing literature, other relevant works include [6] and [7], where compressive sensing is unified with parallel MRI. In [8], a matrix completion framework is proposed for the parallel MRI setting. In [9], [10], [11] and [12], dictionary learning and faster transform learning methods are shown to provide considerably better quality of reconstructions compared to nonadaptive methods. In [13], [14], [15] and [16], low rank models are used for improved results in dynamic MRI setting, for which dictionary-based approaches are also presented in [17] and [18]. Another notable work in the dynamic setting is [19], which, in a compressive sensing framework, generalizes the previous work that exploits spatiotemporal correlations for improved frame rate [20]. Patient-adaptive methods for dynamic MRI also exist in the literature [21], [22]. A Bayesian approach is taken in [23] and [24] for compressive MRI applications, whereas in [25], Bayesian and dictionary learning approaches are combined.

In [26], a deep convolutional network is used to learn the aliasing artefacts, providing a more accurate reconstruction in the case of uniform sampling. In [27], this approach is applied to the radial acquisition setting. In [28], a deep network is used to train the transformations and parameters present in an regularized objective function. Moreover, in [29], a framework based on generative adversarial networks is applied for an improved compressive MRI performance, whereas in [30], a convolutional network is trained for faster acquisition and reconstruction for dynamic MRI setting.

B. Subsampling pattern design

Generally speaking, the most popular approaches to designing $\Omega$ for compressive MRI make use of random variable-density sampling according to a non-uniform probability distribution [31]. The random sampling is done in a manner that favors taking more samples at low frequencies. Some examples include variable-density Poisson disk sampling [32], multi-level sampling schemes [33], [34], pseudo-2D random sampling [35], and variable density with continuous and block sampling models [36–38].

While such variable-density approaches often perform well, they have notable limitations. First, they typically require parameters to be tuned (e.g., the rate of decay of probability away from the center). Second, it is generally unclear which particular sampling distribution will be most effective for a given decoding rule and anatomy. Finally, the very idea of randomization is questionable, since in practice one would like to design a fixed sampling pattern to use across many subjects.

Recently, alternative design methods have been proposed that make use of fully sampled training data (i.e., training signals). In [39]–[41], the training data is used to construct a sampling distribution, from which the samples are then drawn randomly. In [42], [43], a single training image is used at a time to choose a row to sample, and in [44] the rows are chosen based on a mutual information criterion. Much like the above-mentioned randomized variable-density sampling approaches, these existing adaptive algorithms contain parameters for their mask selection whose tuning is non-trivial. Moreover, to our knowledge, none of these works have provided theoretical justifications of the mask selection method. On the other hand, except for [43], these algorithms do not optimize the sampling pattern for a given general decoder. We achieve this via a parallelizable greedy algorithm that implements the given decoder on multiple training images at each iteration of the algorithm until the desired rate is attained.

A particularly relevant prior work is that of [45], in which we proposed the initial learning-based framework that motivates the present paper. However, the focus in [45] is on a simple linear decoder and the noiseless setting, and the crucial aspects of non-linear decoding and noise were left as open problems.

An alternative approach to optimizing subsampling based on prior information is given in [46]. The idea therein is that if a subject requires multiple similar scans, then the previous scans can be used to adjust both the sampling and the decoding of future scans. This is done using the randomized variable-density approach, with the probabilities adjusted to favor locations where the previous scans had more energy. In [47], a proposed informative random sampling approach
optimizes the sampling of subsequent frames of dynamic MRI data based on previous frames in real time. In [48], a highly undersampled pre-scan is used to learn the energy distribution of the image and design the sampling prior to the main scan.

A recent comparative study [49] showed that the approaches that directly use training data perform better than the purely parametric (e.g., randomized variable-density) methods.

Other subsampling designs work include the following: In [50], a generalized Rosetta shaped sampling pattern is used for compressive MRI, and in [51] a random-like trajectory based on higher order chirp sequences is proposed. Radial acquisition designs have been proposed to improve the performance of compressive MRI in the settings of dynamic MRI [52] and phase contrast MRI [53]. In addition, a recent work [54] considered non-Cartesian trajectory design for high resolution MRI imaging at 7T (Tesla).

C. Theory of compressive sensing

The theory of CS has generally moved in very different directions to the above practical approaches. In particular, when it comes to subsampled Fourier measurements, the vast majority of the literature has focused on guarantees for recovering sparse signals with uniform random sampling [55], which performs very poorly in practical imaging applications.

A recent work [33] proposed an alternative theory of CS based on *sparsity in levels*, along with variable-density random sampling and BP decoding. As we outline below, we adopt an entirely different approach that avoids making *any* specific structural assumptions, yet can exploit even richer structures beyond sparsity and its variants.

D. Our contributions

In this paper, we propose a novel learning-based framework for designing subsampling patterns, based on the idea of directly maximizing the empirical performance on training data. We adopt an entirely different theoretical viewpoint to that of the existing CS literature; rather than placing structural assumptions (e.g., sparsity) on the underlying signal, we simply think of the training and test signals as coming from a common unknown distribution. Using connections with statistical learning theory, we adopt a learning method that automatically extracts the structure inherent in the signal, and optimizes $\Omega$ specifically for the decoder at hand.

While our framework is suited to general CS scenarios, we focus on the application of MRI, in which we observe several advantages over the above existing approaches:

- While our previous work [45] exclusively considered a simple linear decoder, in this paper we consider targeted optimization for general non-linear decoders;
- We present a non-trivial extension of our theory and methodology to the noisy setting, whereas [45] only considered the noiseless case;
- We directly optimize for the performance measure at hand (e.g., PSNR), as opposed to less direct measures such as mutual information. Similarly, our framework permits the direct optimization of bottom-line costs (e.g., acquisition time), rather than auxiliary cost measures (e.g., number of samples);
- We can directly incorporate practical sampling constraints, such as the requirement of sampling entire rows and/or columns rather than arbitrary patterns;
- Parameter tuning is not required;
- Our learning algorithm is highly parallelizable, rendering it feasible even when the dimension and/or the number of training images is large;
- We demonstrate the effectiveness of our approach on several real-world data sets, performing favorably against existing methods.

III. Learning-based Framework

A. Overview

Our learning-based framework is outlined as follows:

- We have access to a set of fully-sampled training signals $x_1, \ldots, x_m$ that are assumed to be representative of the unknown signal of interest $x$.
- We assume that the decoder (2) is given. This decoder is allowed to be arbitrary, meaning that our framework can be used alongside general existing reconstruction methods, and potentially also future methods.
- For any subsampling pattern $\Omega$, we can consider its empirical average performance on the training signals:

$$\frac{1}{m} \sum_{j=1}^{m} \eta_{\Omega}(x_j),$$

where $\eta_{\Omega}(x)$ is a performance measure (e.g., PSNR) associated with the signal $x$ and its reconstruction when the sampling pattern is $\Omega$. If $x_1, \ldots, x_m$ are similar to $x$, we should expect that any $\Omega$ such that (5) is high will also perform well on $x$.

- While maximizing (5) is computationally challenging in general, we can use any preferred method to seek an approximate maximizer. We will pay particular attention to a greedy algorithm, which is parameter-free and satisfies a useful nestedness property.

We proceed by describing these points in more detail. For convenience, we initially consider the noiseless setting,

$$b = P_\Omega \Psi x,$$

and then turn to the noisy setting in Section III-F.

B. Preliminaries

Our broad goal is to determine a good subsampling pattern $\Omega \subseteq \{1, \ldots, p\}$ for compressive MRI. To perform this task, we assume that we have access to a set of training signals $x_1, \ldots, x_m$, with $x_j \in \mathbb{C}^p$. The idea is that if an unseen signal has similar properties to the training signals, then we should expect the learned subsampling patterns to generalize well.

In addition to the reconstruction rule $g$ of the form (2), the learning procedure has knowledge of a performance measure, which we would like to make as high as possible on the unseen signal. We focus primarily on PSNR in our experimental section, while also considering the SSIM index.
For implementation reasons, one may wish to restrict the sampling patterns in some way, e.g., to contain only horizontal and/or vertical lines. To account for such constraints, we assume that there exists a set $S$ of subsets of $\{1, \ldots, p\}$ such that the final sampling pattern must take the form

$$\Omega = \bigcup_{j=1}^{\ell} S_j, \quad S_j \in S$$

(7)

for some $\ell > 0$. If $S = \{\{1\}, \ldots, \{p\}\}$, then we recover the setting of [45] where the subsampling pattern may be arbitrary. However, arbitrary sampling patterns are not always feasible; for instance, masks consisting of only horizontal and/or vertical lines are often considered much more suited to practical implementation, and hence, it may be of interest to restrict $S$ accordingly.

Finally, we assume there exists a cost function $c(\Omega) \geq 0$ associated with each subsampling pattern, and that the final cost must satisfy

$$c(\Omega) \leq \Gamma$$

(8)

for some $\Gamma > 0$. We will focus primarily on the case that the cost is the total number of indices in $\Omega$ (i.e., we are placing a constraint on the sampling rate), but in practical scenarios one may wish to consider the ultimate underlying cost, such as the scan time. We assume that $c(\cdot)$ is monotone with respect to inclusion, i.e., if $\Omega_1 \subseteq \Omega_2$ then $c(\Omega_1) \leq c(\Omega_2)$.

C. Theoretical motivation via statistical learning theory

Before describing our main algorithm, we present a theoretical motivation for our learning-based framework. To do so, we think of the underlying signal of interest $x$ as coming from a probability distribution $P$. Under any such distribution, we can write down the indices with the best average performance:

$$\Omega^* = \arg \max_{\Omega \in \mathcal{A}} \mathbb{E}_{P} [\eta_\Omega(x)],$$

(9)

where $\mathcal{A}$ is the set of feasible $\Omega$ according to $c(\cdot)$, $\Gamma$, and $S$, and we define

$$\eta_\Omega(x) = \eta(x, \hat{x})$$

(10)

with $\hat{x} = g(\Omega, b)$ and $b = P_{\Omega} \Psi x$.

Unfortunately, the rule in (9) is not feasible in practice, since one cannot expect to know $P$ (e.g., one cannot reasonably form an accurate probability distribution that describes a brain image). However, if the training signals $x_1, \ldots, x_m$ are also independently drawn from $P$, then there is hope that the empirical average is a good approximation of the true average. This leads to the following selection rule:

$$\hat{\Omega} = \arg \max_{\Omega \in \mathcal{A}} \frac{1}{m} \sum_{j=1}^{m} \eta_\Omega(x_j).$$

(11)

The rule (11) is an instance of empirical risk minimization in statistical learning theory. While finding the exact maximum can still be computationally hard, this viewpoint will nevertheless dictate that we should seek indices $\Omega \in \mathcal{A}$ such that $\frac{1}{m} \sum_{j=1}^{m} \eta_\Omega(x_j)$ is high.

To see this more formally, we consider the following question: If we find a set of indices $\Omega \in \mathcal{A}$ with a good empirical performance $\frac{1}{m} \sum_{j=1}^{m} \eta_\Omega(x_j)$, does it also provide good performance $\mathbb{E}[\eta_\Omega(x)]$ on an unseen signal $x$? The following proposition answers this question in the affirmative using statistical learning theory.

**Proposition 1.** Consider the above setup with a performance measure normalized so that $\eta(x, \hat{x}) \in [0, 1]$. For any $\delta \in (0, 1)$, with probability at least $1 - \delta$ (with respect to the randomness of $x_1, x_2, \ldots, x_m$), it holds that

$$\left| \frac{1}{m} \sum_{j=1}^{m} \eta_\Omega(x_j) - \mathbb{E}_P [\eta_\Omega(x)] \right| \leq \sqrt{\frac{1}{2m} \log \left( \frac{2|\mathcal{A}|}{\delta} \right)},$$

simultaneously for all $\Omega \in \mathcal{A}$.

The proof is given in the appendix. We see that as long as $m$ is sufficiently large compared to $|\mathcal{A}|$, the average performance attained by any given $\Omega \in \mathcal{A}$ on the training data is an accurate estimate of the true performance. This guarantee is with respect to the worst case, regarding all possible probability distributions $P$; the actual performance could exceed this guarantee in practice.

D. Greedy algorithm

While finding the exact maximizer in (11) is challenging in general, we can seek to efficiently find an approximate solution. There are several possible ways to do this, and Proposition 1 reveals that regardless of how we come across a mask with better empirical performance, we should favor it. In this subsection, we present a simple greedy approach, which is parameter-free in the sense that no parameter tuning is needed for the mask selection process once the decoder is given (though the decoder itself may still have tunable parameters).

The greedy approach also exhibits a useful nestedness property (described below).

At each iteration, the greedy procedure runs the decoder $g$ with each element of $S$ that is not yet included in the mask, and adds the subset $S \in \mathcal{S}$ that increases the performance function most on average over the training images, normalized by the cost. The algorithm stops when it is no longer possible to add new subsets from $S$ without violating the cost constraint. The details are given in Algorithm 1.

An important feature of this method is the nestedness property that allows one to immediately adapt for different costs $\Gamma$ (e.g., different sampling rates). Specifically, one can record the order in which the elements are included in $\Omega$ during the mask optimization for a high cost, and use this to infer the mask corresponding to lower costs, or to use as a starting point for higher costs. Note that this is not possible for most parametric methods, where changing the sampling rate requires one to redo the parameter tuning.

We briefly note that alternative greedy methods could easily be used. For instance:

As a concrete example, suppose we are interested in the squared error $\|x - \hat{x}\|_2^2$. If the input is normalized to $\|x\|_2^2 = 1$, then it can be shown that any estimate $\hat{x}$ only improves if it is scaled down such that $\|\hat{x}\|_2^2 \leq 1$. It then follows easily that $\eta(x, \hat{x}) = 1 - \frac{1}{4}\|x - \hat{x}\|_2^2$ always lies in $[0, 1]$. 


Algorithms 1 Greedy mask optimization

**Input:** Training data \( x_1, \ldots, x_m \), reconstruction rule \( g \), sampling subsets \( \mathcal{S} \), cost function \( c \), maximum cost \( \Gamma \)

**Output:** Sampling pattern \( \Omega \)

1: \( \Omega \leftarrow \emptyset \);
2: while \( c(\Omega) \leq \Gamma \) do
3: for \( S \in \mathcal{S} \) such that \( c(\Omega \cup S) \leq \Gamma \) do
4: \( \Omega' = \Omega \cup S \);
5: For each \( j \), set \( b_j \leftarrow P_{\Omega'} \Psi x_j, \hat{x}_j \leftarrow g(\Omega', b_j) \);
6: \( \eta(\Omega') \leftarrow \frac{1}{m} \sum_{j=1}^{m} \eta(x_j, \hat{x}_j) \);
7: \( \Omega \leftarrow \Omega \cup S^* \), where \( S^* = \underset{S : c(\Omega \cup S) \leq \Gamma}{\text{arg max}} \frac{\eta(\Omega \cup S) - \eta(\Omega)}{c(\Omega \cup S) - c(\Omega)} \);
8: return \( \Omega \)

- One could start with \( \Omega = \{1, \ldots, p\} \) (i.e., sampling the entire Fourier space) and then remove samples until a feasible pattern is attained;
- One could adopt a hybrid approach in which samples are both added and removed iteratively until some convergence condition is met.

In our experiments, however, we focus in the procedure in Algorithm 1, which we found to work well.

In another related work [43], an iterative approach is taken in which only a single nonlinear reconstruction is implemented in each iteration of mask selection, starting with an initial mask whereas we run separate reconstructions for each candidate to be added to the sampling pattern, starting with the empty set. Moreover, [43] makes use of several parameters, such as the number of the regions with higher errors to which the samples are moved iteratively, the size of these regions, the power of the polynomial used for a weighting function, etc., which need to be tuned for each experiment. Our greedy algorithm has the advantage of avoiding such heuristics and additional parameters. While the proposed algorithm requires a larger number of computations for mask selection, these computations can be easily parallelized and performed efficiently.

**E. Parametric approach with learning**

An alternative approach is to generate a number of candidate masks \( \Omega_1, \ldots, \Omega_L \), using one or more parametric variable-density methods (possibly with a variety of different choices of parameters), and then to apply the learning-based idea to these candidate masks: Choose the one with the best empirical performance on the training set. While similar ideas have already been used when performing parameter sweeps in existing works (e.g., see [39]), our framework provides a more formal justification to why the empirical performance is the correct quantity to optimize. The details are given in Algorithm 2, where we assume that all candidate masks are feasible according to the sampling subsets \( \mathcal{S} \) and cost function \( c \).

**F. Noisy setting**

So far, we have considered the case that both the acquired signal \( b \) and the training signals \( x_1, \ldots, x_m \) are noiseless. In this subsection, we consider a noisy variant of our setting: The acquired signal is given by

\[
\mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x} + \mathbf{w}
\]

for some noise term \( \mathbf{w} \in \mathbb{C}^p \), and the learning algorithm does not have access to the exact training signals \( x_1, \ldots, x_m \), but instead to noisy versions \( z_1, \ldots, z_m \), where

\[
z_j = x_j + v_j, \quad j = 1, \ldots, m
\]

with \( v_j \) representing the noise.

We observe that the selection rule in (11) can no longer be used, since the learning algorithm does not have direct access to the training signal in the choice of \( \Omega_j \). It turns out, however, that we can do better if we have access to a denoiser \( \xi(z) \) that reduces the noise level. Specifically, suppose that

\[
\xi(z) = x_j + \tilde{v}_j
\]

for some reduced noise \( \tilde{v}_j \) such that \( \mathbb{E}[\|\tilde{v}_j\|] \leq \mathbb{E}[\|v_j\|] \). We then propose the selection rule

\[
\hat{\Omega} = \underset{\Omega \in \mathcal{A}}{\text{arg max}} \frac{1}{m} \sum_{j=1}^{m} \eta(x_j + \tilde{v}_j, \hat{x}(\mathbf{P}_\Omega \Psi (x_j + v_j))),
\]

where \( \hat{x}(\mathbf{b}) \) denotes the decoder applied to \( \mathbf{b} \). Note that we still use the noisy training signal in the choice of \( \mathbf{b} \); by doing so, we are learning how to denoise, which is necessary because the unseen test signal is itself noisy as per (12).

To understand how well the above rule generalized to unseen signals, we would like to compare the empirical performance on the right-hand side of (15) to the true average performance on an unseen signal, defined as

\[
\eta_{\text{noisy}}(\Omega) = \mathbb{E}_x[\eta(x, \hat{x})]
\]

with \( \hat{x} = g(\Omega, \mathbf{b}) \) and \( \mathbf{b} = \mathbf{P}_\Omega \Psi \mathbf{x} + \mathbf{w} \). The following proposition quantifies this comparison.

**Proposition 2.** Consider the above noisy setup with \( \mathbf{w} \) and \( \{v_j\}_{j=1}^{m} \) having independent Gaussian entries of the same variance, and a performance measure \( \eta(x, \hat{x}) \in [0, 1] \) that satisfies the continuity assumption \( |\eta(x, \hat{x}) - \eta(x', \hat{x}')| \leq L\|x - x'\|_2 \) for all \( x, x' \) and some \( L > 0 \). For any \( \delta \in (0, 1) \),
with probability at least $1 - \delta$ (with respect to the randomness of the noisy training signals), it holds that
\[
\left| \frac{1}{m} \sum_{j=1}^{m} \eta(x_j + \tilde{v}_j, \hat{x}(P_{\Omega} \Psi(x_j + v_j)) - \tilde{f}_{\text{noisy}}(\Omega) \right| \\
\leq L \mathbb{E}[\|v\|_2] + \sqrt{\frac{1}{2m} \log \left( \frac{2|\Omega|}{\delta} \right)}.
\]

simultaneously for all $\Omega \in \mathcal{A}$, where $\tilde{v}_j = \xi(v_j)$ is the effective noise remaining in the $j$-th denoised training signal, and $\tilde{v}$ has the same distribution as any given $\tilde{v}_j$.

The proof is given in the appendix. We observe that the second term coincides with that of the noiseless case in Proposition 1, whereas the first term represents the additional error due to the residual noise after denoising. It is straightforward to show that such a term is unavoidable in general.\(^5\)

Hence, along with the fact that more training signals leads to better generalization, Proposition 2 reveals the intuitive fact that the ability to better denoise the training signals leads to better generalization. In particular, if we can do perfect denoising (i.e., $\|\tilde{v}_j\| = 0$) then we get the same generalization error as the noiseless case.

In Algorithm 3, we provide the learning-based procedure with an arbitrary denoising function $\xi$. Note that if we choose the identity function $\xi(z) = z$, then we reduce to the case where no denoising is done.

Algorithm 3 Learning-based mask selection with denoising

**Input**: Noisy training data $z_1, \ldots, z_m$, reconstruction rule $g$, denoising algorithm $\xi(z)$, and either the triplet $(S, c, \Gamma)$ or candidate masks $\Omega_1, \ldots, \Omega_L$.

**Output**: Sampling pattern $\Omega$.

1: $x_j' \leftarrow \xi(z_j)$ for $j = 1, \ldots, m$.
2: Select $\Omega$ using Algorithm 1 or 2 with $\eta(x_j', \hat{x}(P_{\Omega} \Psi z_j))$ replacing $\eta(x_j, \hat{x}(P_{\Omega} \Psi x_j))$ throughout.
3: return $\Omega$.

IV. NUMERICAL EXPERIMENTS

In this section, we provide numerical experiments demonstrating that our learning-based framework provides high-performing sampling patterns for a diverse range of reconstruction algorithms. Our simulation code and data are publicly available online.\(^4\)

A. Implementation details

**Reconstruction rules.** We consider the decoders described in Section II-A, which we refer to as BP, TV, BM3D, and NN (i.e., neural network). For BP in (3), we let the sparsifying operator $\Phi$ be the shearlet transform [56], and for both BP and TV, we implement the minimization using NESTA [57], for which we set the maximum number iterations to 20000, the denoising parameter to $\epsilon = 0$, the tolerance value and the smoothing parameter to $\mu = 10^{-5}$, and the number of continuation steps to $T = 1$.

For BM3D, we use the code available in [58]. We take the observation fidelity parameter $\alpha = 0$, the number of outer iterations $\mathcal{J} = 20$ and the regularization parameters as $\lambda_{\text{max}} = 200$ and $\lambda_{\text{min}} = 0.01$. We also use a varying number of inner iterations between 1 and 10 as described in [3].

For the NN decoder, we use the network structure from [5], only slightly modifying certain parameters. We choose depth of the architecture as $n_d = 3$ and depth of the cascade as $n_c = 5$. We set the mini-batch size for training to 20. We use the same training signals for learning indices and tuning the network weights. Since it is difficult to optimize these jointly, we perform alternating optimization: Initialize the weights, and then alternate between learning indices with fixed weights, and learning weights with fixed indices. We perform up to three iterations of this procedure, which we found to be sufficient for convergence.

As was done in [5], we initialize the network weights using the initialization of He et al. [59], and perform network weight optimization using the Adam algorithm [60] with step size $\alpha = 10^{-2}$ and decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Moreover, we apply an additional $\ell_2$ weight regularization penalty of $10^{-6}$. Each time we train the network, we run the training for 7000 epochs (i.e., passes over the training data). We use the Python implementation available in [5].

In principle, it may sometimes be preferable to change the reconstruction parameters as the greedy algorithm adds indices and increases the current sampling rate. However, we did not find such an approach to provide further benefit in the present setting, so here we stick to the above approach where the reconstruction parameters remain fixed.

**Mask selection methods.** In addition to the greedy method in Algorithm 1, we consider parametric randomized variable-density methods with learning-based optimization according to Algorithm 2; the details are provided in the relevant subsections below. Moreover, we consider the following two baselines from the existing literature:

- (Coherence-based) We consider the parametric approach of [31] with parameters specifying (i) the size of a fully-sampled region at low frequencies; and (ii) the polynomial rate of decay of sampling at higher frequencies. As suggested in [31], we choose the parameters to optimize an incoherence function, meaning that no training data is used. The minimization is done using Monte Carlo methods, and we do this using the code used in [31] available online.

- (Single-image) We consider the approach of [41] in which only a single training image is used. Specifically, this image determines a probability density function where the probability is proportional to energy, and then the samples are randomly selected by drawing from this distribution.

**Data sets.** The MRI data used in the following subsections was acquired on a 3T MRI system (Magnetom Trio Scanner, Erlangen, Germany). The protocols were approved by the...
local ethics committee, and all subjects gave written informed consent.

The data set used in the first three experiments (subsections) below consists of 2D T1-weighted brain scans of seven healthy subjects, which were scanned with a FLASH pulse sequence and a 12-channel receive-only head coil. In our experiments, we use 20 slices of sizes 256×256 from five such subjects (two for training, three for testing). Data from individual coils was processed via a complex linear combination, where coil sensitivities were estimated from an 8×8 central calibration region of k-space [61]. The acquisition used a field of view (FOV) of 220×220 mm² and a resolution of 0.9×0.7 mm². The slice thickness was 4.0 mm. The imaging protocol comprised a flip angle of 70°, a TR/TE of 250.0/2.46 ms, with a scan time of 2 minutes and 10 seconds.

The data set used in subsection E below consists of angiographic brain scans of five healthy subjects acquired with 12-channel receive-only head coil and 20 slices from each are used in our experiments (two subjects for training, three subjects for testing). The size of the slices is 256×256. A 3D TOF sequence was used with FOV of 204×204×51 mm³, 0.8×0.8×0.8 mm³ resolution, flip angle of 18°, magnetization-transfer contrast, a TR/TE of 47/4.6 ms, and a scan time of 16 min 25 sec.

**B. Comparison to baselines**

We first compare to the above-mentioned baselines for a single specific decoder, namely, BP. We use a conventional method of sampling in which readouts are performed as lines at different phase encodes, corresponding to a horizontal line in Fourier space. Hence, our subsampling masks consist of only full horizontal lines, and we let S in Section III-B be the set of all horizontal lines accordingly.

We use our greedy algorithm to find a subset of such lines at a given budget on the total number of samples (or equivalently, the sampling rate). From the data of the five subjects with 20 slices each, we take the first 2 subjects (40 slices total) as training data. Once the masks are obtained, we implement the reconstructions on the remaining 3 subjects (60 slices total). As seen in Figure 1, the learning-based approach outperforms the baselines across all sampling rates shown.

**C. Cross-performances of decoders**

Next continuing as the previous subsection, we compare all four decoders (TV, BP, BM3D, and NN), and evaluate how a mask optimized for one decoder performs when applied to a different decoder. We refer to these as cross-performances, and the results are shown in Table I. Here we report both the PSNR (top) and SSIM values (bottom), but the training optimizes only the PSNR; see Section IV-E for training with respect to the SSIM.

Once again, the learning-based approach outperforms the baselines by approximately 2.5-3.5 dB for all decoders considered. We observe that the greedy method always finds the best performing mask for the given reconstruction algorithms that we use. The performance drop is typically small when the masks optimized for other decoders are used. In Figure 2, we further illustrate these observations with a single slice from the test data, showing the masks and reconstructions along with their PSNR and SSIM values. In this figure, we also compare to a pure low-pass mask given in the bottom row. It can be seen that the greedy masks outperform the the low pass mask as well, in terms of PSNR, SSIM and also visual quality, as they offer sharper images with less aliasing artefacts by balancing between low and high frequency components. On the other hand, as can be seen from the zoomed-in regions, the pure low-pass mask introduces strong blurring, whereas the other baseline masks (coherence-based and single image) cause highly visible aliasing due to suboptimal sampling across low to intermediate frequencies.

![Fig. 1. PSNR as a function of subsampling rates with BP reconstruction.](image-url)
with 256 CPU nodes: (TV) 2 hours and 41 minutes; (BP with shearlets) 3 hours and 23 minutes; (BM3D) 5 hours and 24 minutes. The coherence-based algorithm takes 10 seconds on 256 nodes. The single-image based adaptive algorithm is
Table II: PSNR and SSIM Performances at 25% Subsampling Rate

<table>
<thead>
<tr>
<th>Mask</th>
<th>TV-PSNR</th>
<th>TV-SSIM</th>
<th>BP-PSNR</th>
<th>BP-SSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>parametric</td>
<td>35.02</td>
<td>0.913</td>
<td>36.00</td>
<td>0.926</td>
</tr>
<tr>
<td>greedy</td>
<td>35.89</td>
<td>0.927</td>
<td>37.37</td>
<td>0.942</td>
</tr>
</tbody>
</table>

As seen in Table II, the greedy approach outperforms the parametric approach for both the TV and BP reconstruction algorithms. Interestingly, the masks obtained are also visually rather different (cf., Figures 3 and 4, which also show the reconstructions for a single slice), with the greedy masks being more “spread” rather than taking a continuum of rows at low frequencies. It can be also noticed that both methods choose more horizontal lines than vertical lines, due to the fact the energy in $k$-space is distributed relatively more broadly across the horizontal direction, as can be seen on the top-right corner of Figure 2.

D. Comparison of greedy and parametric methods

We now perform an experiment comparing the greedy approach (Algorithm 1) and the parametric approach with learning (Algorithm 2). In contrast with the previous experiments, we consider measurements in the 2D Fourier space along both horizontal and vertical lines. As described in [35], this is done via a pulse sequence program that switches between phase encoding and frequency encoding, and can provide improvements over the approach of using only horizontal lines.

We tune the parameters of [35] on the training data using Algorithm 2. The first two of the three parameters are $d_x$ and $d_y$, which are the sizes of the fully sampled central regions in horizontal and vertical directions. We sweep this for $d_x, d_y \in \{2, 4, \ldots, d_{\text{max}}\}$ where $d_{\text{max}}$ is maximum feasible fully sampled region size for a given subsampling rate. The last parameter $D$ is the degree of the polynomial that defines the probability distribution function from which random masks are drawn. We sweep over $D \in \{1, 3, 5, \ldots, 13\}$. We then randomly draw 5 masks for each choice of parameters, and we use the mask that gives the best average PSNR on the training data, as per Algorithm 2.

E. Cross-performances of performance measures

In the previous experiments, we focused on the PSNR performance measure. Here we show that considering different measures can lead to different optimized masks, and that it is important to learn a pattern targeted to the correct performance measure. Specifically, we consider both the PSNR and the structural similarity index (SSIM) [1]. Also different from the previous experiments, we use the data set of angiographic brain scans instead of T1-weighted scans (see Section IV-A for details). We return to the method of taking horizontal lines only in the sampling pattern.

Table III gives the PSNR and SSIM performances for the TV and BP decoders, under the masks obtained via the greedy algorithm (cf., Algorithm 1) with the two different performance measures and the decoders at 30% sampling rate. These results highlight the fact that certain decoders are often better suited to certain performance measures. Here, TV is suited to the PSNR measure, as both tend to prefer concentrating the sampling pattern at low frequencies, whereas BP is better suited to SSIM, with both preferring a relatively higher proportion of high frequencies. Note also that in some columns, the performance is not highest on the rows where the training is matched to the decoder and the performance measure (shown in bold), but slightly lower than the highest
TABLE III

RECONSTRUCTION PERFORMANCES AT 30% SUBSAMPLING RATE
AVERAGED OVER 60 ANGIO TEST SLICES. THE CASES THAT THE TRAINING
IS MATCHED TO THE PERFORMANCE MEASURE AND DECODER ARE
HIGHLIGHTED IN BOLD.

<table>
<thead>
<tr>
<th>Decoder</th>
<th>Metric</th>
<th>TV</th>
<th>BP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM</td>
<td>PSNR</td>
<td>SSIM</td>
</tr>
<tr>
<td>Coherence-based</td>
<td>0.738</td>
<td>29.82</td>
<td>0.698</td>
</tr>
<tr>
<td>Single-Image</td>
<td>0.766</td>
<td>30.72</td>
<td>0.744</td>
</tr>
<tr>
<td>SSIM-greedy-TV</td>
<td>0.795</td>
<td>31.90</td>
<td>0.779</td>
</tr>
<tr>
<td>PSNR-greedy-TV</td>
<td>0.799</td>
<td>32.74</td>
<td>0.806</td>
</tr>
<tr>
<td>SSIM-greedy-BP</td>
<td>0.797</td>
<td>32.72</td>
<td>0.806</td>
</tr>
<tr>
<td>PSNR-greedy-BP</td>
<td>0.795</td>
<td>32.54</td>
<td>0.804</td>
</tr>
<tr>
<td>Low Pass</td>
<td>0.771</td>
<td>31.70</td>
<td>0.777</td>
</tr>
</tbody>
</table>

values, which is most likely either due to limited training data
or the suboptimality of the greedy algorithm.

These observations are further illustrated in Figure 5 (only
for TV decoder and its masks due to space constraints), where
we show the optimized masks, the reconstructions on a single
slice and as the maximum intensity projection (MIP) of the
volume this slice belongs to [62]. We see in particular that
the two masks are somewhat different, with that for the PSNR
containing more gaps at higher frequencies and fewer gaps at
lower frequencies. We also observe that compared to the data
used in the previous subsections, the angiographic data used
in this experiment is more concentrated at the center of the
$k-$space. The greedy algorithm is able to adapt to this change
and obtain masks that have more lower frequencies.

F. Experiments with additive noise

The data we used in the previous subsections has very low
levels of noise. In order to test the validity of our claims in the
noisy setting, we add bivariate circularly symmetric complex
random Gaussian noise to our normalized complex images,
with a noise standard deviation of $\sigma = 3 \times 10^{-4}$ for both the
real and imaginary components. Since the ground truth images
are normalized, this noise level gives an average signal-to-
noise ratio (SNR) of 25.68 dB. We set the denoising parameter
of NESTA to $\epsilon = 10$ for TV minimization and to $\epsilon = 1.1$ for
BP with shearlets which work well with the various masks
and images used in this section. In Algorithm 1, we measure
the error at each iteration with respect to denoised image that
is obtained using BM3D denoising algorithm [4]. Note that
the ground truth should not be used in the learning algorithm,
since it is unknown in practice. On the other hand, in the
testing part, we compute the errors with respect to the ground
truth images.

As can be seen from Figure 6 and Table IV, the greedy
algorithm is still capable of finding a better mask compared
to the other baseline masks. Therefore, in this example, our
approach is robust with respect to noise. Note that we train
with respect to the PSNR, but also report the SSIM values.
Note also that compared to the case where the noise levels
were very low, the mask obtained in noisy setting is slightly
closer to a low-pass mask. The reason for this is that the noise
hides the relatively weaker signal present at high frequencies,
while only having a minimal effect on the stronger signal
present at lower frequencies.


Table IV

<table>
<thead>
<tr>
<th>Decoder</th>
<th>TV</th>
<th></th>
<th>BP</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SSIM</td>
<td>PSNR</td>
<td>SSIM</td>
<td>PSNR</td>
</tr>
<tr>
<td>Coherence-based</td>
<td>0.799</td>
<td>29.73</td>
<td>0.808</td>
<td>30.29</td>
</tr>
<tr>
<td>Single-Image</td>
<td>0.828</td>
<td>30.86</td>
<td>0.843</td>
<td>31.58</td>
</tr>
<tr>
<td>TV-greedy</td>
<td>0.878</td>
<td>33.07</td>
<td>0.896</td>
<td>34.21</td>
</tr>
<tr>
<td>BP-greedy</td>
<td>0.875</td>
<td>32.89</td>
<td>0.893</td>
<td>34.07</td>
</tr>
<tr>
<td>Low Pass</td>
<td>0.859</td>
<td>31.37</td>
<td>0.872</td>
<td>31.80</td>
</tr>
</tbody>
</table>

V. Conclusion

We have presented a versatile learning-based framework for selecting masks for compressive MRI, using training signals to optimize for a given decoder and anatomy. As well as having a rigorous justification via statistical learning theory, our approach is seen to provide improved performance on real-world data sets for a variety of reconstruction methods. Since our framework is suited to general decoders, it can potentially be used to optimize the indices for new reconstruction methods that are yet to be discovered. In this work, we focused on 1D subsampling for 2D MRI, 2D subsampling (via horizontal and vertical lines) for 2D MRI, and 1D subsampling for 3D MRI, but our greedy approach can potentially provide an automatic way to optimize the sampling in the settings of 2D subsampling for 3D MRI and non-Cartesian sampling, as opposed to constructing a randomized pattern on a case-by-case basis. For the setting of 3D MRI, there is an additional computational challenge to our greedy algorithm, since the candidate set is large.

In future studies, we will also seek to validate the performance under the important practical variation of multi-coil measurements, as well as applications beyond MRI such as computer tomography, phase retrieval, and ultrasound. We finally note that in this paper, the number of subjects and training images used was relatively small, and we anticipate that larger data sets would be of additional benefit in realizing the full power of our theory.

Appendix

A. Proof of Proposition 1

Using the fact that \( \eta \) lies in \([0, 1]\) and applying Hoeffding’s inequality [63], we obtain for any \( \Omega \in \mathcal{A} \) and \( t > 0 \) that

\[
\left| \frac{1}{m} \sum_{j=1}^{m} \eta_{\Omega}(x_{j}) - \mathbb{E}_{\Phi}[\eta_{\Omega}(x)] \right| \leq t,
\]

with probability at least \( 1 - 2 \exp(-2mt^{2}) \). Since the probability of a union of events is upper bounded by the sum of the individual probabilities (i.e., the union bound), we find that the same inequality holds for all \( \Omega \in \mathcal{A} \) with probability at least \( 1 - 2 |\mathcal{A}| \exp(-2mt^{2}) \). The proposition follows by setting \( \delta = 2 |\mathcal{A}| \exp(-2mt^{2}) \) and solving for \( t \).
Proposition 2

By the fact that the Fourier transform is a unitary operation and i.i.d. Gaussian vectors are invariant under unitary transformations, we have

$$\eta_{\text{noisy}}(\Omega) = E[\eta(x, \hat{x}(P\Omega\Psi x + w))] = E[\eta(x, \hat{x}(P\Omega\Psi(x + v)))],$$

where $$\hat{x}(b)$$ denotes the estimator applied to the noisy output $$b$$, and $$v$$ has the same distribution as any given $$v_j$$.

Let $$\hat{v} = \xi(v)$$ be the denoised version of $$v$$. Using the triangle inequality, we write

$$\left| \frac{1}{m} \sum_{j=1}^{m} \eta(x_j + \hat{v}_j, \hat{x}(P\Omega\Psi(x_j + v_j))) - \eta_{\text{noisy}}(\Omega) \right|$$

$$= \left| \frac{1}{m} \sum_{j=1}^{m} \eta(x_j + \hat{v}_j, \hat{x}(P\Omega\Psi(x_j + v_j))) 
- E[\eta(x, \hat{x}(P\Omega\Psi(x + v)))] \right|$$

$$\leq \left| \frac{1}{m} \sum_{j=1}^{m} \eta(x_j + \hat{v}_j, \hat{x}(P\Omega\Psi(x_j + v_j))) 
- E[\eta(x + \hat{v}, \hat{x}(P\Omega\Psi(x + v)))] \right|$$

$$+ \left| E[\eta(x + \hat{v}, \hat{x}(P\Omega\Psi(x + w)))] - E[\eta(x, \hat{x}(P\Omega\Psi x + w))] \right|$$

Using (18) and following the proof of Proposition 1, the first term is upper bounded by $$\sqrt{\frac{1}{m^2} \log \left( \frac{2A^4}{\delta} \right)}$$ with probability at least $$1 - \delta$$. Moreover, by the continuity condition assumed in the proposition statement, the second term above is upper bounded by $$L \max_{\|v\|}$$, thus completing the proof.

Acknowledgment

This work has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement n° 725594 - time-data), and from Hasler Foundation Program: Cyber Human Systems (project number 16066). It was also sponsored by the Department of the Navy, Office of Naval Research (ONR) under a grant number N62909-17-1-2111.

References


