Learning-Based Compressive Subsampling

Luca Baldassarre, Yen-Huan Li, Jonathan Scarlett, Baran Gözcü, Ilija Bogunovic and Volkan Cevher
Laboratory for Information and Inference Systems
École Polytechnique Fédérale de Lausanne, Switzerland

The problem of recovering a structured signal $\mathbf{x} \in \mathbb{C}^p$ from a set of dimensionality-reduced linear measurements $\mathbf{b} = \mathbf{Ax}$ arises in a variety of applications, such as medical imaging, spectroscopy, Fourier optics, and computerized tomography. Due to computational and storage complexity or physical constraints imposed by the problem, the measurement matrix $\mathbf{A} \in \mathbb{C}^{n \times p}$ is often of the form $\mathbf{A} = \mathbf{P}_\Omega \mathbf{\Psi}$ for some orthonormal basis matrix $\mathbf{\Psi} \in \mathbb{C}^{p \times p}$ and subsampling operator $\mathbf{P}_\Omega : \mathbb{C}^p \to \mathbb{C}^n$ that selects the rows indexed by Ω. This raises the fundamental question of how best to choose the index set Ω in order to optimize the recovery performance. Previous approaches to addressing this question rely on non-uniform *random* subsampling using application-specific knowledge of the structure of $\mathbf{x}$. In this paper, we instead take a principled learning-based approach in which a *fixed* index set is chosen based on a set of training signals $\mathbf{x}_1, \dots, \mathbf{x}_m$. We formulate combinatorial optimization problems seeking to maximize the energy captured in these signals in an average-case or worst-case sense, and we show that these can be efficiently solved either exactly or approximately via the identification of modularity and submodularity structures. We provide both deterministic and statistical theoretical guarantees showing how the resulting measurement matrices perform on signals differing from the training signals, and we provide numerical examples showing our approach to be effective on a variety of data sets.

Index Terms—Compressive sensing, learning-based measurement design, data-driven sensing design, non-uniform subsampling, structured sparsity, scientific and medical imaging, submodular optimization

I. Introduction

In the past decade, there has been a tremendous amount of research on the problem of recovering a structured signal $\mathbf{x} \in \mathbb{C}^p$ (e.g., a signal with few non-zero coefficients) from a set of dimensionality-reduced measurements of the form

$$\mathbf{b} = \mathbf{Ax} + \mathbf{w}, \tag{1}$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a known measurement matrix, and $\mathbf{w} \in \mathbb{C}^n$ represents unknown additive noise.

This problem is often referred to as compressive sensing (CS), and is ubiquitous in signal processing, having an extensive range of applications including medical resonance imaging (MRI), spectroscopy, radar, Fourier optics, and computerized tomography [1]. The key challenges in CS revolve around the identification of structure in the signal (e.g., sparsity), computationally efficient algorithms, and measurement matrix designs. In this paper, we focus primarily on the latter.

In the classical CS setting where $\mathbf{x}$ is sparse, the best theoretical recovery guarantees are obtained by random matrices with independent and identically distributed (iid) entries,

along with sparsity-promoting convex optimization methods [2]. However, such random matrices impose a significant burden in the data processing pipeline in terms of both storage and computation as the problem dimension p increases.

To overcome these limitations, significant attention has been paid to subsampled structured matrices of the form [2]

$$\mathbf{A} = \mathbf{P}_\Omega \mathbf{\Psi}, \tag{2}$$

where $\mathbf{\Psi} \in \mathbb{C}^{p \times p}$ is an orthonormal basis, and $\mathbf{P}_\Omega : \mathbb{C}^p \to \mathbb{C}^n$, where $[\mathbf{P}_\Omega \mathbf{x}]_l = \mathbf{x}_{\Omega_l}$, $l = 1, \dots, n$, is a subsampling operator that selects the rows indexed by the set Ω, with $|\Omega| = n$. Popular choices for $\mathbf{\Psi}$ include Fourier and Hadamard matrices, both of which permit efficient storage and computation.

Measurements of the form (2) are also of direct interest in applications, such as MRI and Fourier optics, where one cannot choose the basis in which measurements are taken, but one does have the freedom in choosing a subsampling pattern. Constraining the measurement matrix in this manner raises the following fundamental question: *How do we best choose the index set Ω?* The answer depends on the structure known to be inherent in $\mathbf{x}$, as well as the recovery algorithm used.

For general sparse vectors, uniform random subsampling provides a variety of recovery guarantees under suitable assumptions on the incoherence of $\mathbf{\Psi}$ with the sparsity basis of the signal to be recovered [2]. However, it has been observed that this approach can be significantly outperformed in applications via *non-uniform* random subsampling techniques, in which one takes more samples in the parts of the signal that tend to contain more energy, thus exploiting structured sparsity in the signal as opposed to sparsity alone [3]. Explicit distributions on Ω are proposed in [3], [4], containing parameters that require tuning for the application at hand.

From a practical perspective, the randomization of Ω is typically not desirable, and perhaps not even possible. Moreover, there is strong evidence that in real-world applications there is a "best" choice of Ω that one should aim for. For instance, while the compression standard [5] does not require it, the subsampling for JPEG in the discrete cosine transform (DCT) domain is typically chosen independently of the image being compressed, in order to simplify the encoder. The indices for MRI systems, which are based on Fourier measurements, are often randomly chosen based on matching image statistics and performing manual tuning, and are then fixed for later scans [3], [4], [6], [7]. Despite its importance, principled approaches to the "best" index selection problem appear to be lacking.

To bridge this gap, our paper takes a *learning-based* approach to subsampling via training signals and combinatorial optimization problems. The high-level idea is simple: We

select the indices that preserve as much energy as possible in a set of training signals, either in a worst-case or average-case sense, and we show that this is also equivalent to minimizing the ℓ_2-error achieved by a simple linear decoder. By identifying combinatorial optimization structures such as submodularity, we show that we can find exact or near-exact solutions to these optimization problems in polynomial time. We then provide both deterministic and statistical theoretical guarantees characterizing how well the selected set of indices perform when applied to new signals differing from the training set. Finally, we demonstrate the effectiveness of our approach on a variety of data sets, showing matching or improved performance compared to [3], [4], [6] in several imaging applications.

Both our theory and our recovery results are based on the use of highly efficient linear encoder *and decoder* pairs, and we provide some examples that challenge the conventional wisdom in CS that non-linear decoding methods, such as basis pursuit, are necessary for reliable recovery [8]. Indeed, we find that although such methods are needed when considering *arbitrary* sparse or compressible signals, we may avoid them in certain settings where the signals and the sampling mechanisms exhibit more specific structures. Nevertheless, non-linear extensions of our work would also be of considerable interest; see Section V for further discussion.

Our main theoretical contribution is our statistical generalization bound, which states that when the training and test signals are independently drawn from *any* common distribution, the average ℓ_2-norm error of the above-mentioned linear decoder for the test signal can be made arbitrarily close to the best possible, including randomized strategies, when the number of training signals exceeds a value depending on the complexity of the set constraints, e.g., $O\left(n \log \frac{p}{n}\right)$ training signals suffice when there are no constraints.

A. Problem Statement

For clarity of exposition, we focus on the noiseless case throughout the paper, and consider the measurement model

$$\mathbf{b} = \mathbf{P}_\Omega \boldsymbol{\Psi} \mathbf{x} \qquad (3)$$

for some orthonormal basis matrix $\boldsymbol{\Psi} \in \mathbb{C}^{p \times p}$, and some subsampling matrix $\mathbf{P}_\Omega$ whose rows are canonical basis vectors.

Given a set of m training signals $\mathbf{x}_1, \ldots, \mathbf{x}_m$, we seek to select an index set Ω permitting the recovery of further signals that are in some sense similar to the training signals; see Section III for some formal notions of similarity. We assume without loss of generality that $\|\mathbf{x}_j\|_2 = 1$ for all j. While our focus is on the selection of Ω, it is clear that the recovery algorithm also plays a role towards achieving our goal.

Linear Decoding: The main procedure that we consider simply expands $\mathbf{b}$ to a p-dimensional vector by placing zeros in the entries corresponding to Ω^c, and then applies the adjoint $\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^{-1}$:

$$\hat{\mathbf{x}} = \boldsymbol{\Psi}^* \mathbf{P}_\Omega^T \mathbf{b}. \qquad (4)$$

Note that this is a *linear* decoder, and can be implemented highly efficiently even in large-scale systems for suitably structured matrices $\boldsymbol{\Psi}$ (e.g., Fourier or Hadamard). In fact, it

is easily shown to be equivalent to the *least-squares* decoder, i.e., the pseudo-inverse of $\mathbf{P}_\Omega \boldsymbol{\Psi}$ is $\boldsymbol{\Psi}^* \mathbf{P}_\Omega^T$ for unitary $\boldsymbol{\Psi}$.

Basis Pursuit Decoding: If $\mathbf{x}$ is known to be approximately sparse in some known basis, i.e., $\mathbf{x} = \boldsymbol{\Phi}^* \mathbf{z}$ for some approximately sparse vector $\mathbf{z}$ and basis matrix $\boldsymbol{\Phi}$, then stable and robust recovery is possible using standard CS algorithms. A particularly popular choice is basis pursuit (BP), which estimates

$$\hat{\mathbf{z}} = \underset{\tilde{\mathbf{z}} \,:\, \mathbf{b} = \mathbf{P}_\Omega \boldsymbol{\Psi} \boldsymbol{\Phi}^* \tilde{\mathbf{z}}}{\arg\min} \|\tilde{\mathbf{z}}\|_1 \qquad (5)$$

and then sets $\hat{\mathbf{x}} = \boldsymbol{\Phi}^* \hat{\mathbf{z}}$. We can also replace the basis pursuit recovery with other convex programs that leverage additional "structured" sparsity of the coefficients, e.g., see [9].

We will use both decoders (4) and (5) in our numerical results in Section IV, and in fact see that they behave similarly in all of the examples therein, despite BP having a significantly higher computational complexity.

B. Related Work

Variable-density Subsampling: To the best of our knowledge, all previous works on variable-density subsampling have considered randomized choices of Ω and sought the corresponding distributions, rather than considering fixed choices. A common approach is to design such distributions using empirically-observed phenomena in images [3], [4] or adopting specific signal models such as generalized Gaussian [11]. In each of these works, the proposed designs involve parameters that need to be learned or tuned, and we are not aware of any efficient principled approaches for doing so. An alternative approach is taken in [12] based on minimizing coherence (see also [13] for a related work with unstructured matrices), but the optimization is only based on the measurement and sparsity bases, as opposed to training data.

Learning-based Measurement Designs: Several previous works have proposed methods for designing *unstructured* measurement matrices based on training data, with a particularly common approach being to seek the best restricted isometry property (RIP) constant [2] with respect to the training data. In [14], this problem is cast as an affine rank minimization problem, which is then relaxed to a semidefinite program. A different approach is taken in [15], where the affine rank minimization problem is reformulated and approximated using game-theoretic tools. An alternating minimization approach is adopted in [16] in the context of metric learning. A drawback of all of these works is that since the measurement matrices are unstructured, they suffer from similar storage and computation limitations to those of random matrices with iid entries.

C. Contributions

The main contributions of this paper are as follows:

- We present a class of optimization problems for selecting the subsampling indices based on training data, seeking to choose those that capture as much energy as possible with respect to the average case or worst case, or equivalently, that minimize the ℓ_2 error achieved by a simple linear decoder (see Section II-A). For each of these, we identify inherent modularity and submodularity structures, thus

permitting exact or approximate solutions to be obtained efficiently using existing algorithms.

- For the average-case criterion, we show that a simple sorting procedure achieves the exact solution. Moreover, we show that several additional constraints on the subsampling pattern can be incorporated while still permitting an efficient solution, including matroid constraints [17] and wavelet-tree based constraints [18].
- We provide theoretical results stating how the selected indices perform when applied to a test signal not present in the training set. For both the average-case and worst-case criteria, this is done from a deterministic perspective, assuming the new signal to be sufficiently close to the training data in a sense dictated by Ω. For the average-case criterion, we also provide bounds from a statistical perspective, assuming the data to be independently drawn from an unknown probability distribution, and drawing connections with empirical risk minimization.
- We demonstrate the effectiveness of our approach on a variety of data sets, in particular showing matching or improved performance compared to [3], [4], [11] in certain imaging scenarios.

D. Organization of the Paper

In Section II, we formally introduce our optimization problems, our techniques for solving them, and the resulting optimality guarantees. Our theoretical results are presented in Section III, namely, the deterministic and statistical generalization bounds. In Section IV, we demonstrate the effectiveness of our approach on a variety of data sets, and provide comparisons to previous approaches. Conclusions are drawn in Section V.

II. LEARNING-BASED SUBSAMPLING STRATEGIES

In this section, we motivate our learning-based approach to subsampling, formulate the corresponding optimization problems in terms of average-case and worst-case criteria, and propose computationally efficient algorithms for solving them exactly or approximately.

A. Motivation

The idea behind our subsampling strategy is simple: Given the training signals $\mathbf{x}_1, \ldots, \mathbf{x}_m$, we seek a subsampling scheme that preserves as much of their energy as possible. Perhaps the most intuitively simple optimization problem demonstrating this idea is the following:

$$\hat{\Omega} = \arg\max_{\Omega \,:\, |\Omega|=n} \min_{j=1,\ldots,m} \|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}_j\|_2^2. \qquad (6)$$

Beyond the natural interpretation of capturing energy, this can also be viewed as optimizing the worst-case performance of the linear decoder proposed in (4) with respect to the error in the ℓ_2-norm. Indeed, substituting (3) into (4), we obtain

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = \|\mathbf{x} - \mathbf{\Psi}^* \mathbf{P}_\Omega^T \mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}\|_2^2 \qquad (7)$$

$$= \|\mathbf{\Psi}\mathbf{x} - \mathbf{P}_\Omega^T \mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}\|_2^2 \qquad (8)$$

$$= \|\mathbf{P}_{\Omega^c}^T \mathbf{P}_{\Omega^c} \mathbf{\Psi} \mathbf{x}\|_2^2 \qquad (9)$$

$$= \|\mathbf{P}_{\Omega^c} \mathbf{\Psi} \mathbf{x}\|_2^2 \qquad (10)$$

where (8) follows since $\mathbf{\Psi}$ is an orthonormal basis matrix and thus $\mathbf{\Psi}\mathbf{\Psi}^* = \mathbf{I}$, (9) follows since $\mathbf{P}_\Omega^T \mathbf{P}_\Omega + \mathbf{P}_{\Omega^c}^T \mathbf{P}_{\Omega^c} = \mathbf{I}$, and (10) follows since a multiplication by $\mathbf{P}_{\Omega^c}^T$ simply produces additional rows that are equal to zero. By decomposing the energy according to the entries on Ω and Ω^c, we have

$$\|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}\|_2^2 + \|\mathbf{P}_{\Omega^c} \mathbf{\Psi} \mathbf{x}\|_2^2 = 1. \qquad (11)$$

whenever $\|\mathbf{x}\|_2 = 1$, and substitution into (10) yields

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 = 1 - \|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}\|_2^2. \qquad (12)$$

Thus, maximizing the objective in (6) amounts to minimizing the ℓ_2-norm error for the decoder in (4).

Recalling that $\mathbf{\Psi}$ is an orthonormal basis matrix, we have $\|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}_j\|_2 \leq \|\mathbf{x}_j\|_2$. Thus, defining $\mathbf{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ and $\mathbf{V} := \mathbf{\Psi}\mathbf{X}$, we can equivalently write (6) as

$$\hat{\Omega} = \arg\min_{\Omega \,:\, |\Omega|=n} \|\mathbf{1} - \mathrm{diag}(\mathbf{V}^T \mathbf{P}_\Omega^T \mathbf{P}_\Omega \mathbf{V})\|_\infty, \qquad (13)$$

where $\mathbf{1}$ is the vector of m ones, and $\mathrm{diag}(\cdot)$ forms a vector by taking the diagonal entries of a matrix. In this form, the optimization problem can also be interpreted as finding the subsampling pattern providing the best restricted isometry property (RIP) constant [2] with respect to the training data, analogously to the optimization problems of [15], [16] for unstructured (rather than subsampled) matrices.

B. Optimization Criteria and Constraints

Generalizing (6), we study the following class of problems:

$$\hat{\Omega} = \arg\max_{\Omega \in \mathcal{A}} F(\Omega), \qquad (14)$$

where $\mathcal{A} \subseteq \{\Omega \,:\, |\Omega| = n\}$ is a cardinality constrained subset of $\{1, \ldots, p\}$, and the set function $F(\cdot)$ is given by

$$F(\Omega) := f\big(\|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}_1\|_2^2, \ldots, \|\mathbf{P}_\Omega \mathbf{\Psi} \mathbf{x}_m\|_2^2\big) \qquad (15)$$

for some function f. For example, (6) is recovered by setting $f(\alpha_1, \ldots, \alpha_m) = \min_{j=1,\ldots,m} \alpha_j$ and $\mathcal{A} = \{\Omega \,:\, |\Omega| = n\}$. While this choice of $\mathcal{A}$ is perhaps the most obvious, one may be interested in more restrictive choices imposing structured constraints on the subsampling pattern; some examples are given below.

The optimization problem in (14) is combinatorial, and in general finding the exact solution is NP hard (e.g., this is true for the special case in (6); see [19]). The key idea in all of the special cases below is to identify advantageous combinatorial structures in the problem in order to efficiently obtain near-optimal solutions. In particular, we will see that *submodularity* structures in the objective function, and *matroid* structures in the constraint set, play a key role. We proceed by presenting these definitions formally.

Definition 1. *A set function $h(\Omega)$ mapping subsets $\Omega \subseteq \{1, \ldots, p\}$ to real numbers is said to be* submodular *if, for all $\Omega_1, \Omega_2 \subset \{1, \ldots, p\}$ with $\Omega_1 \subseteq \Omega_2$, and all $i \in \{1, \ldots, p\} \backslash \Omega_2$, we have*

$$h(\Omega_1 \cup \{i\}) - h(\Omega_1) \geq h(\Omega_2 \cup \{i\}) - h(\Omega_2). \qquad (16)$$

The function is said to be modular *if the same holds true with equality in place of the inequality.*

This definition formalizes the notion of diminishing returns: Adding an element to a smaller set increases the objective function more compared to when it is added to a larger set. Our focus in this paper will be on submodular functions that are also *monotone*, i.e., $h(\Omega_2) \geq h(\Omega_1)$ whenever $\Omega_1 \subseteq \Omega_2$.

Submodular or modular set functions often allow us to efficiently obtain near-optimal solutions with matroid constraints [17] which we now define.

Definition 2. *Given $V = \{1, \ldots, n\}$ and a non-empty set of subsets (of V) $\mathcal{A}$, the pair $(V, \mathcal{A})$ is said to be a matroid if (i) for any $A \in \mathcal{A}$, all subsets $A' \subseteq A$ are also in $\mathcal{A}$; (ii) given $A \in \mathcal{A}$ and $B \in \mathcal{A}$ with $|B| > |A|$, there exists an element $v \in B \backslash A$ such that $A \cup \{v\} \in \mathcal{A}$.*

We will see in Section III-B that the cardinality of the constraint set $\mathcal{A}$ plays a key role in determining the number of training signals needed to obtain statistical guarantees on near-optimality.

In the context of structured signal recovery, a particularly notable example of a matroid constraint is *multi-level subsampling* [6], where the indices $\{1, \ldots, p\}$ are split into K disjoint groups with sizes $\{p_k\}_{k=1}^K$, and the number of measurements within the k-th group is constrained to be n_k, with $\sum_{k=1}^K n_k = n$. This corresponds to a matroid known as the *partition matroid* [17]. In this case, the total number of possible sparsity patterns is $\prod_{k=1}^K \binom{p_k}{n_k}$. As opposed to the *random* multi-level subsampling scheme in [6], our framework also optimizes the samples within each level given the sparsity constraints.

We can also go beyond matroid constraints; we mention one additional example here. In the context of image compression with image-independent subsampling, $\mathbf{\Psi}$ may correspond to the wavelet basis, and a suitable choice for $\mathcal{A}$ forces the coefficients to form a rooted connected subtree of the wavelet tree of cardinality n [18], [20]. In this case, the total number of subsampling patterns is the Catalan number, $\frac{1}{n+1}\binom{2n}{n}$ [18]. This does not correspond to a matroid constraint, but it can nevertheless be handled using dynamic programming [20].

C. Average-case Criterion ($f = f_{\mathrm{avg}}$)

We first consider the function $f_{\mathrm{avg}}(\alpha_1, \ldots, \alpha_m) := \frac{1}{m}\sum_{j=1}^m \alpha_j$, yielding the optimization problem

$$\hat{\Omega} = \arg\max_{\Omega \in \mathcal{A}} \frac{1}{m} \sum_{j=1}^m \sum_{i \in \Omega} |\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|^2, \qquad (17)$$

where $\boldsymbol{\psi}_i$ is the transpose of the i-th row of $\mathbf{\Psi}$. This corresponds to maximizing the *average* energy in the training signals, which may be preferable to the worst-case criterion in (6) due to an improved robustness to outliers, e.g., corrupted training signals.

Since the sum of (sub)modular functions is again (sub)modular [21], we see that (17) is a modular maximization problem, thus permitting an exact solution to be found efficiently in several cases of interest.

Case 1 (No Additional Constraints): In the case that $\mathcal{A} = \{\Omega : |\Omega| = n\}$, the exact solution is found by sorting: *Select the n indices whose values of $\frac{1}{m}\sum_{j=1}^m |\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|^2$*

are the largest. The running time is dominated by the pre-computation of the values $\frac{1}{m}\sum_{j=1}^m |\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|^2$, and behaves as $O(mp^2)$ for general matrices $\mathbf{\Psi}$, or $O(mp \log p)$ for suitably structured matrices such as Fourier and Hadamard.

Case 2 (Matroid Constraints): In the case that $\mathcal{A}$ corresponds to a matroid, we can also find the exact solution by a simple greedy algorithm [17]: *Start with the empty set, and repeatedly add the item that increases the objective value by the largest amount without violating the constraint, terminating once n indices have been selected.* The values $|\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|^2$ can be computed in $O(mp^2)$ (general case) or $O(mp \log p)$ time (structured case), and the greedy algorithm itself can be implemented in $O(nmp)$ time, with the factor of m arising due to the summation in (17).

In some cases, the complexity can be further improved; in particular, for the above-mentioned multi-level sampling constraint, the exact solution is found by simply performing sorting within each level.

Case 3 (Other Constraints): As hinted above, the types of constraints that can be efficiently handled in modular optimization problems are not limited to matroids. Choosing the set $\mathcal{A}$ that forces the coefficients to form a rooted connected wavelet tree [18], [20], there exists a dynamic program for finding the optimal solution in $O(nmp)$ time [20]. We can also obtain exact solutions for *totally unimodular* constraints via linear programming (LP) relaxations; see [9] for an overview in the context of sparse recovery.

D. Generalized Average-case Criterion ($f = f_{\mathrm{gen}}$)

We generalize the choice $f = f_{\mathrm{avg}}$ by considering $f_{\mathrm{gen}}(\alpha_1, \ldots, \alpha_m) := \frac{1}{m}\sum_{j=1}^m g(\alpha_j)$, yielding

$$\hat{\Omega} = \arg\max_{\Omega \in \mathcal{A}} \frac{1}{m} \sum_{j=1}^m g\left(\sum_{i \in \Omega} |\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|^2 \right), \qquad (18)$$

for some function $g : [0, 1] \to \mathbb{R}$. We consider the case that g is an increasing concave function with $g(0) = 0$.

Note that our framework also permits weighted averages of the form $\frac{1}{\sum_j \beta_j} \sum_j \beta_j g(\alpha_j)$, which may be relevant when some training signals are more reliable than others. More generally, we can allow for weights β_{ij} on the terms $|\langle \boldsymbol{\psi}_i, \mathbf{x}_j \rangle|$, which may be desirable when is is more important to capture the energy in certain parts of the signal than in others. In this paper, we focus on uniform weights for clarity of exposition.

By simple rearrangements, we can see that the choice $g(\alpha) = 1 - (1-\alpha)^q$ ($q \geq 1$) in (18) is equivalent to considering (13) with $\|\cdot\|_\infty$ replaced by $\|\cdot\|_q^q$. In particular, the case $q = 1$ recovers (17), and $q = 2$ is another reasonable choice that is in a sense "in between" the choices $q = 1$ and $q = \infty$. Section IV shows that the choice $q = 2$ performs well numerically.

We established above that the argument to g in (18) is a modular function of Ω. Recalling that g is concave and increasing by assumption, it follows that (18) is a submodular optimization problem [21, Sec. 1.1]. While finding the exact solution is hard in general, we can efficiently find an approximate solution with rigorous guarantees in several cases.

Case 1 (No Additional Constraints): In the case that $\mathcal{A} = \{\Omega : |\Omega| = n\}$, a solution whose objective value is within a

multiplicative factor of $1 - \frac{1}{e}$ of the optimum can be found via the greedy algorithm [22].

Case 2 (Matroid Constraints): With matroid constraints, the greedy algorithm is only guaranteed optimality to within a factor of $\frac{1}{2}$ [17]. However, there also exist polynomial-time algorithms for obtaining solutions that are within a factor of $1 - \frac{1}{e}$ of the optimum; for example, see [23], [24] and the references therein.

E. Worst-case Criterion ($f = f_{\min}$)

Finally, we consider the choices $\mathcal{A} = \{\Omega : |\Omega| = n\}$ and $f_{\min}(\alpha_1, \ldots, \alpha_m) := \min_{j=1,\ldots,m} \alpha_j$, yielding

$$\hat{\Omega} = \underset{\Omega \,:\, |\Omega|=n}{\arg\max} \; \min_{j=1,\ldots,m} \sum_{i\in\Omega} |\langle \boldsymbol{\psi}_i, \mathbf{x}_j\rangle|^2, \tag{19}$$

which coincides with (6). This can be thought of as seeking robustness with respect to the "worst" image in the training set, which may be desirable in some cases. To our knowledge, the algorithm that we consider below has not been extended to more general choices of $\mathcal{A}$, and we thus leave such cases for future work.

The objective function in (19) is the minimum of m modular functions. This form of optimization problem was studied in [19], where an algorithm called Saturate was proposed, which has guarantees for the general template

$$\max_{\Omega \,:\, |\Omega|\leq n} \; \min_{j=1,\ldots,m} \; f_j(\Omega), \tag{20}$$

where $f_1, \cdots, f_m$ are monotone submodular functions with $f_j(\emptyset) = 0$. The algorithm takes a parameter α representing how much the set size may exceed the "target" value n, and a parameter ϵ used as a stopping criterion. The details are shown in Algorithm 1, and the greedy partial cover (GPC) subroutine is shown in Algorithm 2.

Algorithm 1 Saturate($f_1, \cdots, f_m, n, \alpha, \epsilon$) [19]
___
1: $c_{\min} \leftarrow 0; c_{\max} \leftarrow \min_j f_j(\{1, \ldots, p\}); \Omega_{\text{best}} \leftarrow \emptyset$
2: **while** $(c_{\max} - c_{\min}) > \epsilon$ **do**
3: $\quad c \leftarrow (c_{\min} + c_{\max})/2$
4: $\quad \bar{f}_c(\Omega) \leftarrow \frac{1}{m} \sum_{j=1}^m \min\{f_j(\Omega), c\}$
5: $\quad \hat{\Omega} \leftarrow \text{GPC}(\bar{f}_c, c)$
6: $\quad$ **if** $|\hat{\Omega}| > \alpha n$ **then**
7: $\quad\quad c_{\max} \leftarrow c$
8: $\quad$ **else**
9: $\quad\quad c_{\min} \leftarrow c; \Omega_{\text{best}} \leftarrow \hat{\Omega}$

Algorithm 2 GPC($\bar{f}_c, c$) (Greedy Partial Cover) [19]
___
1: $\Omega \leftarrow \emptyset$
2: **while** $\bar{f}_c(\Omega) < c$ **do**
3: $\quad \Delta_j \leftarrow \bar{f}_c(\Omega \cup \{j\}) - \bar{f}_c(\Omega)$
4: $\quad \Omega \leftarrow \Omega \cup \arg\max_j \Delta_j$

It is shown in [19, Thm. 5] that for integer-valued functions $\{f_j\}_{j=1}^m$ the Saturate algorithm finds a solution $\hat{\Omega}$ such that

$$\min_j f_j(\hat{\Omega}) \geq \max_{|\Omega|\leq k} \min_j f_j(\Omega), \quad \text{and} \quad |\hat{\Omega}| \leq \alpha k \tag{21}$$

with $\alpha = 1 + \log\big(\max_{i=1,\ldots,p} \sum_{j=1}^m f_j(\{i\})\big)$. While we do not have integer-valued functions in our setting, we may use the observation from [19, Sec. 7.1] that in the general case, analogous guarantees can be obtained at the expense of having an additional term in α depending linearly on the number of bits of accuracy. We observe that when applied to (19), we have $\log\big(\max_{i=1,\ldots,p} \sum_{j=1}^m f_j(\{i\})\big) = \log\big(\max_{i=1,\ldots,p} \sum_{j=1}^m |\langle \boldsymbol{\psi}_i, \mathbf{x}_j\rangle|\big)$, which is in turn upper bounded by $\log m$ since both $\boldsymbol{\psi}_i$ and $\mathbf{x}_j$ have unit norm. Thus, even with this crude bounding technique, this term only constitutes a logarithmic factor.

Moreover, it was observed empirically in [19] that the Saturate algorithm can provide state-of-the-art performance even when the functions are not integer-valued and α is set to one. Thus, although the theory in [19] does not directly capture this, we expect the algorithm to provide a good approximate solution to (19) even without the logarithmic increase in the number of measurements.

The running time is $O(p^2 m \log m)$ in the worst case [19]. In practice, we observe the algorithm to run much faster, as was also observed in [19]. Moreover, we found the total number of samples returned by the algorithm to be very close to its maximum value αn (e.g., within 1%).

III. GENERALIZATION BOUNDS

The optimization problems in the previous section seek to capture as much of the signal energy as possible on the training signals $\mathbf{x}_j$, which also corresponds to minimizing the ℓ_2-norm error of the decoder in (4) (*cf.*, Section II-A). However, it is not immediately clear to what extent the same will be true on a new signal $\mathbf{x}$. In this section, we take two distinct approaches to characterizing this theoretically.

A. Deterministic Approach

This subsection provides simple bounds showing that the energy loss remains small on any new signal $\mathbf{x}$ that is "close" to the training data in a sense dictated by $\Omega^c := \{1, \ldots, p\}\backslash\Omega$. We assume that $\mathbf{x}$ is normalized in the same way as the training signals, i.e., $\|\mathbf{x}\|_2 = 1$; this is without loss of generality, since general signals may be normalized to satisfy this condition, and this normalization constant does not need to be known by the recovery algorithm.

We first consider the generalized average-case criterion, which includes the average-case criterion as a special case.

Theorem 1. (Deterministic generalization bound for $f = f_{\text{gen}}$) *Fix $\delta > 0$ and $\epsilon > 0$, and suppose that for a set of training signals $\mathbf{x}_1, \ldots, \mathbf{x}_m$ with $\|\mathbf{x}_j\|_2 = 1$, we have a sampling set Ω such that*

$$\frac{1}{m} \sum_{j=1}^m g\big(\|\mathbf{P}_\Omega \boldsymbol{\Psi} \mathbf{x}_j\|_2^2\big) \geq g(1 - \delta). \tag{22}$$

Then for any signal $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$ such that $\frac{1}{m} \sum_{j=1}^m \|\mathbf{P}_{\Omega^c} \boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2^2 \leq \epsilon$, we have

$$\|\mathbf{P}_\Omega \boldsymbol{\Psi} \mathbf{x}\|_2^2 \geq 1 - \big(\sqrt{\delta} + \sqrt{\epsilon}\big)^2. \tag{23}$$

Proof. Substituting (11) into (22), we obtain $\frac{1}{m}\sum_{j=1}^{m} g(1 - \|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2) \geq g(1 - \delta)$. Recalling that g is concave by assumption, applying Jensen's inequality on the left-hand side gives $g\left(1 - \frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2\right) \geq g(1 - \delta)$. Since g is also monotonically increasing by assumption, it follows that $\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2 \leq \delta$.

Next, since the average of squares is at least as large as the square of the average, the previous condition implies $\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2 \leq \sqrt{\delta}$, and the assumption $\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2^2 \leq \epsilon$ similarly implies $\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2 \leq \sqrt{\epsilon}$. We thus obtain from the triangle inequality that

$$\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}\|_2$$
$$\leq \frac{1}{m}\sum_{j=1}^{m}\Big(\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2 + \|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2\Big) \quad (24)$$
$$\leq \sqrt{\delta} + \sqrt{\epsilon}. \quad (25)$$

Taking the square and applying (11), we obtain (23). □

The case $f = f_{\text{avg}}$ corresponds to $f = f_{\text{gen}}$ with $g(\alpha) = \alpha$, and in this case, the condition in (22) takes the particularly simple form

$$\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2 \geq 1 - \delta, \quad (26)$$

requiring that at least a fraction $1 - \delta$ of the training signal energy be captured on average.

We now turn to the worst-case criterion.

Theorem 2. (Deterministic generalization bound for $f = f_{\min}$) *Fix $\delta > 0$ and $\epsilon > 0$, and suppose that for a set of training signals $\mathbf{x}_1, \ldots, \mathbf{x}_m$ with $\|\mathbf{x}_j\|_2 = 1$, we have a sampling set Ω such that*

$$\min_{j=1,\ldots,m}\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2 \geq 1 - \delta. \quad (27)$$

Then for any signal $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$ such that $\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2^2 \leq \epsilon$ for some $j \in \{1, \ldots, m\}$, we have

$$\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}\|_2^2 \geq 1 - \left(\sqrt{\delta} + \sqrt{\epsilon}\right)^2. \quad (28)$$

Proof. It follows from (11) and (27) that $\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2 \leq \delta$ for all j. Hence, letting j be an index such that $\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2^2 \leq \epsilon$, we obtain from the triangle inequality that

$$\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}\|_2 \leq \|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}\mathbf{x}_j\|_2 + \|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2 \quad (29)$$
$$\leq \sqrt{\delta} + \sqrt{\epsilon}. \quad (30)$$

Taking the square and applying (11), we obtain (28). □

We note that a sufficient condition for $\|\mathbf{P}_{\Omega^c}\boldsymbol{\Psi}(\mathbf{x} - \mathbf{x}_j)\|_2^2 \leq \epsilon$ is that $\|\mathbf{x} - \mathbf{x}_j\|_2^2 \leq \epsilon$. However, this is certainly not necessary; for example, when $\mathbf{x} = -\mathbf{x}_j$, the latter is large, whereas the former is small provided that the index set Ω captures most of the signal energy.

We note that Theorems 1 and 2 are rather different despite appearing to be similar. In Theorem 2, the definition of δ corresponds to capturing the energy in *all* training signals, and the definition of ϵ corresponds to the new signal being close to *some* training signal (i.e., the minimum distance). In contrast, in Theorem 1, both δ and ϵ are defined with respect to the corresponding average.

B. Statistical Approach

Thus far, we have treated all of our signals as being deterministic. We now turn to a statistical approach, in which the training signals $\mathbf{x}_1, \ldots, \mathbf{x}_m$ and the test signal $\mathbf{x}$ are independently drawn from a common probability distribution $\mathbb{P}$ on $\mathbb{C}^p$. We assume that this distribution is *unknown*, and thus cannot be exploited as prior knowledge for the design of Ω. We also assume that $\|\mathbf{x}\|_2 = 1$ almost surely for $\mathbf{x} \sim \mathbb{P}$; otherwise, we can simply replace $\mathbf{x}$ by $\frac{\mathbf{x}}{\|\mathbf{x}\|_2}$ throughout.

In this setting, there is precise notion of a "best" subsampling set: The set $\Omega^* \in \mathcal{A}$ that captures the highest proportion of the signal energy *on average* is given by

$$\Omega^* = \arg\max_{\Omega \in \mathcal{A}} \mathbb{E}\big[\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}\|_2^2\big]. \quad (31)$$

It is worth noting that, under our statistical model, one cannot improve on this choice by moving to randomized subsampling strategies. To see this, we write for any randomized Ω

$$\mathbb{E}\big[\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}\|_2^2\big] = \mathbb{E}\Big[\mathbb{E}\big[\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}\|_2^2 \,\big|\, \Omega\big]\Big]. \quad (32)$$

Then, by the mean value theorem, there exists a deterministic set such that the argument to the outer expectation has a value at least as high as the average.

Since the distribution $\mathbb{P}$ of $\mathbf{x}$ is unknown, we cannot expect to solve (31) exactly. However, given m training samples, we can approximate the true average by the empirical average:

$$\hat{\Omega} = \hat{\Omega}(\mathbf{x}_1, \ldots, \mathbf{x}_m)$$
$$= \arg\max_{\Omega \in \mathcal{A}} \frac{1}{m}\sum_{j=1}^{m}\|\mathbf{P}_{\Omega}\boldsymbol{\Psi}\mathbf{x}_j\|_2^2. \quad (33)$$

This idea is known as *empirical risk minimization* in statistical learning theory, and in the present setting, we see that it yields precisely the optimization problem in (14) with $f = f_{\text{avg}}$.

We are now interested in determining how the average energy captured by $\hat{\Omega}$ compares to the optimal choice Ω^*. To this end, we introduce some definitions. We let $\mathbb{E}_{\mathbf{x}}[\cdot]$ denote averaging with respect to $\mathbf{x}$ alone, so that for any function $h(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m)$, $\mathbb{E}_{\mathbf{x}}[h(\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m)]$ is a random variable depending on the training signals. With this notation, we define the random variable

$$\Delta_n = \Delta_n(\mathbf{x}_1, \ldots, \mathbf{x}_m) \quad (34)$$
$$:= \mathbb{E}\big[\|\mathbf{P}_{\Omega^*}\boldsymbol{\Psi}\mathbf{x}\|_2^2\big] - \mathbb{E}_{\mathbf{x}}\big[\|\mathbf{P}_{\hat{\Omega}}\boldsymbol{\Psi}\mathbf{x}\|_2^2\big] \quad (35)$$

representing the gap to optimality as a function of the training signals. In the following theorem, we bound this gap *independently of the distribution* $\mathbb{P}$.

Theorem 3. (Statistical generalization bound for $f = f_{\text{avg}}$) *Under the above statistical model, for any $\eta > 0$, we have with probability at least $1 - \eta$ that*

$$\Delta_n \leq \sqrt{\frac{2}{m}\left(\log|\mathcal{A}| + \log\frac{2}{\eta}\right)}. \quad (36)$$

Proof. For brevity, we define $\gamma_\Omega(\mathbf{x}) := \|\mathbf{P}_\Omega \boldsymbol{\Psi} \mathbf{x}\|_2^2$. Moreover, we define the empirical average

$$\hat{\mathbb{E}}_m[\gamma_\Omega(\cdot)] := \frac{1}{m} \sum_{i=1}^m \gamma_\Omega(\mathbf{x}_j), \tag{37}$$

which is a random variable depending on $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

With these definitions, we have the following:

$$\Delta_n = \mathbb{E}[\gamma_{\Omega^*}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}}[\gamma_{\hat{\Omega}}(\mathbf{x})] \tag{38}$$

$$= \left(\mathbb{E}[\gamma_{\Omega^*}(\mathbf{x})] - \hat{\mathbb{E}}_m[\gamma_{\Omega^*}(\cdot)] \right)$$
$$+ \left(\hat{\mathbb{E}}_m[\gamma_{\Omega^*}(\cdot)] - \hat{\mathbb{E}}_m[\gamma_{\hat{\Omega}}(\cdot)] \right)$$
$$+ \left(\hat{\mathbb{E}}_m[\gamma_{\hat{\Omega}}(\cdot)] - \mathbb{E}_{\mathbf{x}}[\gamma_{\hat{\Omega}}(\mathbf{x})] \right) \tag{39}$$

$$\leq \left| \hat{\mathbb{E}}_m[\gamma_{\Omega^*}(\cdot)] - \mathbb{E}[\gamma_{\Omega^*}(\mathbf{x})] \right|$$
$$+ \left| \hat{\mathbb{E}}_m[\gamma_{\hat{\Omega}}(\cdot)] - \mathbb{E}_{\mathbf{x}}[\gamma_{\hat{\Omega}}(\mathbf{x})] \right| \tag{40}$$

$$\leq 2 \max_{\Omega \in \mathcal{A}} \left| \hat{\mathbb{E}}_m[\gamma_\Omega(\cdot)] - \mathbb{E}[\gamma_\Omega(\mathbf{x})] \right|, \tag{41}$$

where (40) follows since $\hat{\mathbb{E}}_m[\gamma_{\Omega^*}(\cdot)] \leq \hat{\mathbb{E}}_m[\gamma_{\hat{\Omega}}(\cdot)]$ by the definition of $\hat{\Omega}$ in (33).

By the assumption that $\|\mathbf{x}\|_2 = 1$, we have $\gamma_\Omega(\mathbf{x}) \in [0, 1]$. Moreover, using the definition in (37) and the fact that both $\mathbf{x}$ and $\{\mathbf{x}_j\}_{j=1}^m$ have distribution $\mathbb{P}$, the average of $\hat{\mathbb{E}}_m[\gamma_\Omega(\cdot)]$ is given by $\mathbb{E}[\gamma_\Omega(\mathbf{x})]$. It thus follows from Hoeffding's inequality [25, Sec. 2.6] that

$$\mathbb{P}\Big[\big| \hat{\mathbb{E}}_m[\gamma_\Omega(\cdot)] - \mathbb{E}[\gamma_\Omega(\mathbf{x})] \big| > t \Big] \leq 2 e^{-2mt^2} \tag{42}$$

for any set Ω and constant $t > 0$. Thus, by the union bound, we have

$$\mathbb{P}\Big[\max_{\Omega \in \mathcal{A}} \big| \hat{\mathbb{E}}_m[\gamma_\Omega(\cdot)] - \mathbb{E}[\gamma_\Omega(\mathbf{x})] \big| > t \Big] \leq 2|\mathcal{A}| e^{-2mt^2}. \tag{43}$$

The proof is concluded by setting $t = \sqrt{\frac{1}{2m} \log(\frac{2}{\eta}|\mathcal{A}|)}$ and substituting the condition of the event into (41). $\qquad\square$

In the case that $\mathcal{A} = \{\Omega : |\Omega| = n\}$, we have $|\mathcal{A}| = \binom{p}{n}$, and hence, Theorem 3 reveals that the performance of $\hat{\Omega}$ can be made arbitrarily close to that of Ω^* using $m = O(n \log \frac{p}{n})$ training signals, for any distribution $\mathbb{P}$. Our numerical findings in Section IV suggest that this bound may be pessimistic in some specific scenarios, as we achieve good results even when m is smaller than n. Nevertheless, it is reassuring that the optimal performance can always be approached when the number of training signals is large enough.

Clearly, considering a smaller set $\mathcal{A}$ yields an improved bound in (36); in particular, this occurs for the multi-level sampling and rooted connected tree structures in accordance with the cardinalities given in Section II-B. On the other hand, using a smaller set can also worsen the performance of Ω^* itself, due to a smaller maximization set in (31). Thus, there is a trade-off between the performance of the best choice within the class considered, and the extent to which the bound guarantees that the corresponding performance is approached using few training samples.

Remark 1. Although we focused on $f = f_{\text{avg}}$ for clarity, the analysis remains valid with $f = f_{\text{gen}}$ provided that g is bounded in $[0, 1]$ (e.g., $g(\alpha) = 1 - (1 - \alpha)^2$). Specifically, we modify (31)–(35) by replacing the squared-norms $\|\cdot\|_2^2$ by $g(\|\cdot\|_2^2)$, and similarly in the definition of $\gamma_\Omega(\mathbf{x})$ in the proof.

IV. NUMERICAL EXPERIMENTS

In this section, we present the results of numerical experiments that illustrate the effectiveness of the learning-based compressive subsampling approach. For a given number of measurements, our consistent observation is that our approach matches or improves the quality of signal recovery over the randomized variable-density based sampling approaches. A likely reason for the improvements is that we directly optimize the sampling indices, as opposed to only optimizing auxiliary parameters that are related to those indices.

We also identify several scenarios in which our learning-based subsampling in tandem with the simple *linear* decoder (4) outperforms randomized sampling techniques used in conjunction with the *non-linear* decoder (5). Moreover, we find in all of our examples that the improvements of the non-linear decoders over the linear one are quite marginal even for the randomized subsampling indices.

This is not conclusive evidence against the use of non-linear decoders in these problems, since it may be possible to devise more sophisticated techniques to optimize Ω specifically for these decoders. Nevertheless, our observations support the practitioners' historical preference of simple decoders.

Throughout this section, the experiments are done using the algorithms proposed in Section II: Sorting for f_{avg}, the greedy algorithm for f_{gen}, and the Saturate algorithm for f_{min}.

A. Kenya & ImageNet Data sets

We created an image data set in 16-bit `tiff` format of 187 Kenya images of resolution 2048×2048 from one of the authors' personal collection, without any compression. Given the high-resolution images, we used MATLAB's `imresize` function to create 1024×1024, 512×512, and 256×256 pixel images. We split the data set into a training set of the first 137 images and a test set with the remaining 50 images. Here are some examples from this data set:

Figure 1 illustrates the compression performance of our index selection using the Hadamard, discrete cosine (DCT), and wavelet (Daub-4) transforms. In the figure, we compute the square root of the empirical average of $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ over the test set for a range of n, and also determine the best adaptive compression (i.e., the best *image-dependent* n-term approximation in the measurement basis) as the baseline, to reach an error level of 0.1, 0.05 and 0.01. Here we normalize so that the total energy in each image is 1.

The learned indices for Hadamard-subsampling and DCT-subsampling get within a factor of 2 of their corresponding adaptive compression rates on the test data. In contrast, the learned indices for wavelets are within a factor of 4 of their adaptive compression rates. While the adaptive DCT and wavelet compressions are close, the learned DCT-subsampling
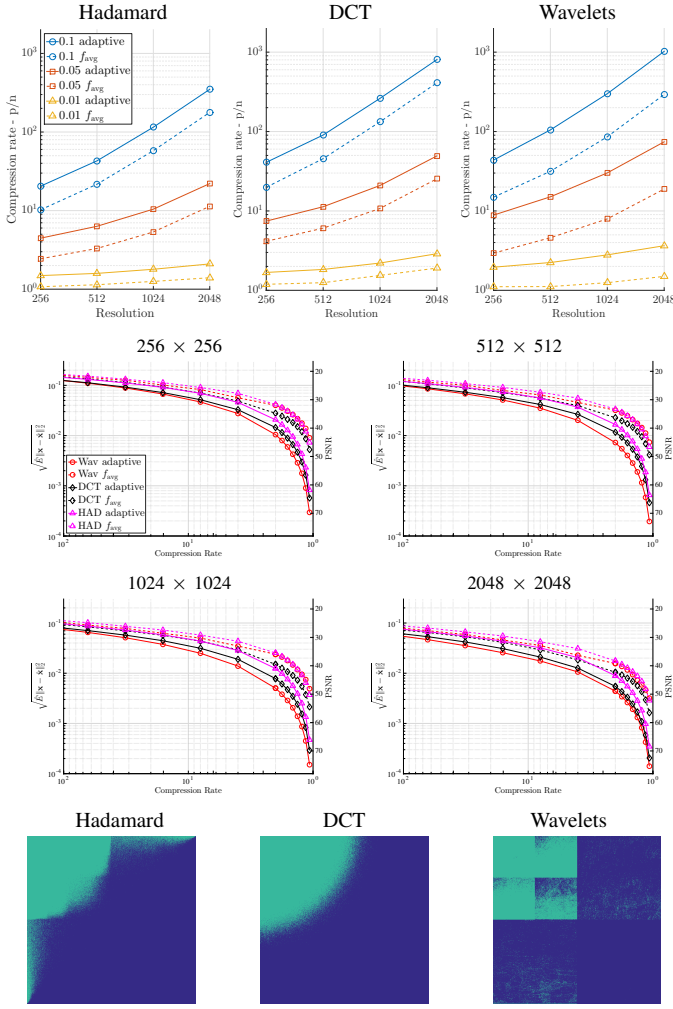
Fig. 1. (Top) Maximum compression factor to achieve a given ℓ_2-error performance. (Middle) Average estimation error in conjunction with average PSNR values. (Bottom) Selected indices at 25% compression with f_{avg}.
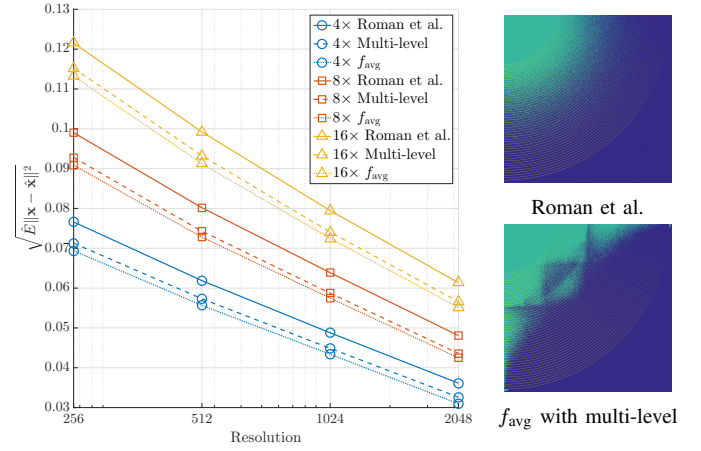

Fig. 2. (Left) Recovery performance with the linear decoder. (Right) Tuned randomized variable-density pattern [Top] vs. Learned indices with multi-level constraints [Bottom]. The overlayed circles are the non-overlapping multi-level partitions.
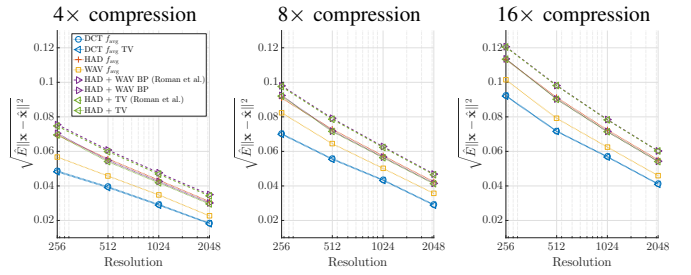

Fig. 3. The performance of the nonlinear decoder vs. the linear decoder.

obtains compression rates within a factor of 2 of the adaptive wavelet compression.

Overall, the learned DCT indices obtain the best compression performance on the test data. Moreover, to quadruple the resolution at an error of 0.1, which corresponds to 25dB peak signal-to-noise ratio (PSNR), we simply need to double the number of samples. While we observe a similar behavior for the average error level of 0.05, we need to commensurately increase the number of samples with resolution for the error level of 0.01, which corresponds to high levels of PSNR values, as shown in Figure 1 (Middle).

Figure 1 (Bottom) shows the learned subsampling patterns. The learned wavelet samples exhibit the expected parent-child correlation on its decomposition hierarchy. Intriguingly, while the learned DCT samples show circular symmetry, the learned Hadamard samples exhibit an unexpected shape which is typically not predicted by randomized approaches.

We made use of indices provided to us by the first author of [4] for Hadamard sampling, which are based on the randomized approach therein but with tuning done for the goal of effectively compressing natural images. See Figure 2 (Right),

where a 25% compression rate is used. Using these indices and the corresponding non-overlapping partitions (also shown in the figure), we counted the number of samples in each partition and created a corresponding set of constraints (*cf.*, the multi-level sampling defined in Section II). We then applied our learning-based approach using the resulting constraint set $\mathcal{A}$. Surprisingly, the shape of the learned pattern differs from the circularly symmetric version and outperforms the randomized indices when we use the linear decoder; see Figure 2 (Left).

The theory associated with the approach of [4] requires the basis pursuit (BP) decoder (5), with the sparsity basis Φ being the wavelet decomposition. Figure 3 illustrates that the overall improvement of this decoder over the linear one is limited, and similarly for an additional non-linear decoder based on minimizing the total variation (TV).
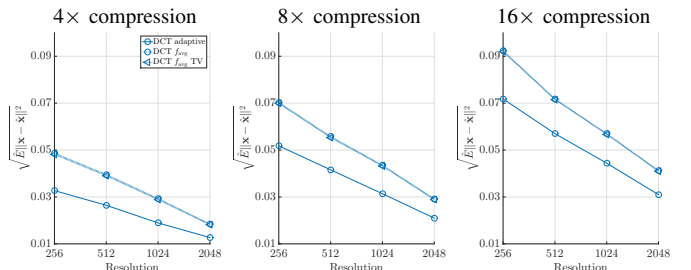

Fig. 4. The performance of the nonlinear decoder vs. the linear decoder for DCT measurements only.

In summary, for this data set, replacing the random index set coming from the randomized approach with the learned indices leads to a noticeable improvement, whereas the improvements from switching to non-linear decoders are marginal. It should be noted that these decoders incur heavy computational costs; some running times are summarized in Table I based on a state-of-the-art primal-dual decomposition method [26].

TABLE I
KENYA IMAGES: ℓ_2-ERRORS VS. RUNNING TIMES

Resolution	Recovery	Sampling rate		
		6.25%	12.50%	25%
256	BP	0.102 / 6s	0.083 / 6s	0.063 / 6s
	TV	0.102 / 27s	0.082 / 22s	0.062 / 20s
	Adjoint	0.103 / 0.01s	0.084 / 0.01s	0.064 / 0.01s
512	BP	0.080 / 23s	0.063 / 22s	0.048 / 22s
	TV	0.080 / 151s	0.063 / 162s	0.047 / 153s
	Adjoint	0.081 / 0.03s	0.064 / 0.03s	0.049 / 0.02s
1024	BP	0.062 / 85s	0.049 / 85s	0.036 / 93s
	TV	0.062 / 340s	0.049 / 614s	0.036 / 65s
	Adjoint	0.063 / 0.08s	0.050 / 0.08s	0.037 / 0.09s
2048	BP	0.047 / 381s	0.036 / 366s	0.026 / 333s
	TV	0.047 / 1561s	0.036 / 2501s	0.025 / 2560s
	Adjoint	0.048 / 0.26s	0.037 / 0.29s	0.027 / 0.28s

As the Kenya data set size is quite limited, we performed similar experiments on a much larger data set called ImageNet.[1] In particular, we consider the 2010 ImageNet Large Scale Visual Recognition Challenge data set. This data set consists of approximately 1.4 million images of various resolutions; see Figure 5 (top) for our training and test splits. Since the data is already JPEG compressed with 8 bits, we crop the existing high-resolution images in the central part of the image in order to obtain low-resolution ones.

While the images are already compressed, Figure 5 (Middle) illustrates that the qualitative behavior we have seen so far does not change: Wavelets achieve the best adaptive compression and the learned DCT subsampling performs the best in the test data. We also observe that the learned indices for the rooted-connected (RC) wavelet tree model generalize marginally better than the learned wavelet sampling. Finally, Figure 5 (Bottom) shows that the distribution of the sampling patterns for Hadamard and DCT do not exhibit circularly symmetric shapes. Moreover, the learned wavelet indices appear to concentrate on the spatial center of the images.

B. iEEG Data set

This subsection focuses on an intracranial electro-encephalography (iEEG) data set from the `iEEG.org` portal.

Currently, iEEG is an invasive procedure that permits the recording of neuro-electrical signals via an array of electrodes placed directly on the brain surface. A very active research area concerns the design of implantable wireless devices that do not require the patient to be tethered to an external recorder. Efficient compression schemes based on compressive sensing have been recently proposed in order to reduce transmission power consumption [27]–[29]; however, they may use a large chip area [28] or not be able to compress beyond the number of channels [29].

We show the effectiveness of our approach on the micro-electrode recordings from the `I001_P034_D01` data set. We consider only the recordings annotated as seizures, with

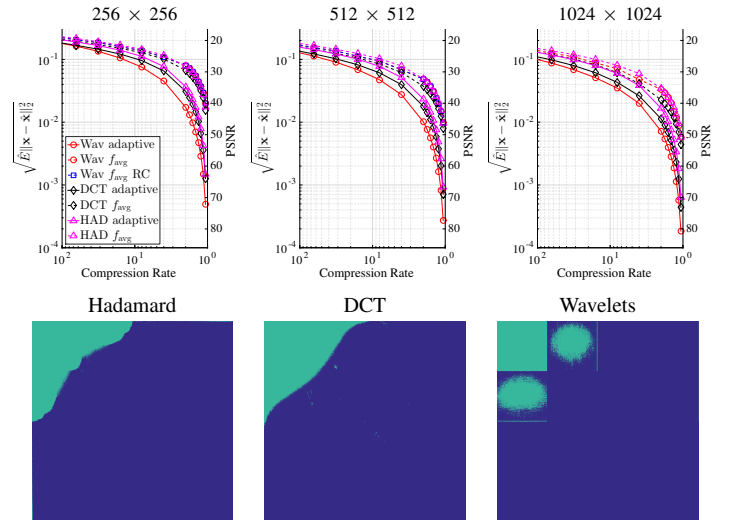[1] Available at http://image-net.org

Fig. 5. (Top) The test and training split sizes on the ImageNet data set. (Middle) Average estimation error in conjunction with average PSNR values. (Bottom) Selected indices at 12.5% compression with f_{avg}.
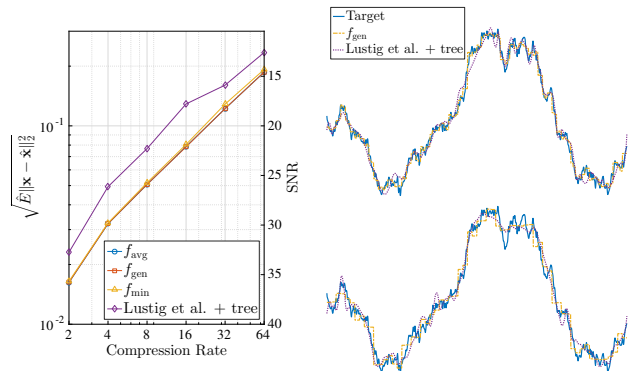

Fig. 6. iEEG data: Recovery performance (Left) and recovery examples (right): $16\times$ compression (Top) and $32\times$ compression (Bottom).

the exception of seizure 7, which is corrupted. Moreover, we remove channel 1, which is inactive, and channel 7, which erroneously records the AC signal powering the system instead of the neuronal signal. We define signal windows of $p = 1024$ samples and use seizures $1, 2, 3, 4, 5, 6, 8$ for learning the indices of the subsampled Hadamard transform for compression, for a total of 5.8 million samples over 5 channels. We test the reconstruction performance of the chosen indices on channels $1, 2, 3, 5$ and 6 on the last seizure. For the randomized variable-density sampling approach, we consider non-linear reconstruction via wavelets and a tree-structure promoting norm, shown in [30] to yield the best performance on this type of data. For our learning-based approach, we again use the simple linear decoder given in (4).

Figure 6 (Left) illustrates that the learning-based approach outperforms the randomized variable-density approach. Note that here we use the function f_{gen} with $g(\alpha) = 1 - (1 - \alpha)^2$, as introduced in Section II-D. We use Hadamard sampling,

TABLE II
OBJECTIVE VALUES OBTAINED FOR VARIOUS RECOVERY CRITERIA ON
THE iEEG DATA SET WITH $p = 1024$. CR STANDS FOR COMPRESSION
RATE.

CR	Criterion	Metric					
		f_{avg}		f_{gen}		f_{min}	
		Train	Test	Train	Test	Train	Test
2×	f_{avg}	0.9980	0.9975	0.999998	0.999997	0.9955	0.9945
	f_{gen}	0.9980	0.9975	0.999998	0.999997	0.9955	0.9945
	f_{min}	0.9977	0.9972	0.999997	0.999996	0.9963	0.9943
4×	f_{avg}	0.9919	0.9902	0.99997	0.99995	0.9843	**0.9804**
	f_{gen}	0.9919	0.9901	0.99997	0.99995	0.9843	**0.9805**
	f_{min}	0.9913	0.9896	0.99996	0.99994	0.9862	**0.9797**
8×	f_{avg}	0.9790	0.9756	0.9998	0.9997	0.9534	0.9539
	f_{gen}	0.9789	0.9757	0.9998	0.9997	0.9555	0.9541
	f_{min}	0.9747	0.9724	0.9997	0.9996	0.9614	0.9504
16×	f_{avg}	0.9466	0.9407	0.9987	0.9982	0.8501	0.8063
	f_{gen}	0.9460	0.9422	0.9987	0.9983	0.8633	0.8338
	f_{min}	0.9372	0.9330	0.9983	0.9977	0.8737	0.8149
32×	f_{avg}	0.8642	0.8618	0.9919	0.9906	0.6734	0.6921
	f_{gen}	0.8642	0.8618	0.9922	0.9906	0.6731	0.6920
	f_{min}	0.8254	0.8299	0.9873	0.9861	0.6970	0.6522
64×	f_{avg}	0.6643	0.6883	0.9554	0.9539	0.4549	0.2939
	f_{gen}	0.6643	0.6883	0.9554	0.9539	0.4550	0.2944
	f_{min}	0.6296	0.6639	0.9462	0.9471	0.4815	0.2946

since it is easy to implement in digital hardware [28]. The density function of [3] is parametrized by the radius of fully sampled region, r, and the polynomial degree, d. We choose the parameters' values that yield the lowest reconstruction error on the training set for each compression rate in the ranges $r = \frac{4}{2p}, \frac{5}{2p}, \ldots, \frac{16}{2p}$ and $d = 1, 2, \ldots, 75$, taking the best realization over 20 random draws in each case.

Figure 6 (Left) shows example reconstructions at $16\times$ and $32\times$ compressions. A likely reason for the reduced error due to our approach is that we do not assume any shape for the distribution of the indices.

In Table II, we present the values of f_{avg}, f_{gen} and f_{min} obtained by our learning-based algorithms for various compression ratios with $p = 1024$. Note that we normalize the signals to have unit energy, and hence 1 is the highest possible value of each objective. As expected, the procedure trained for a given objective always gives the best value of that objective on the training data. However, it is sometimes the case that the indices obtained for optimizing f_{avg} and f_{gen} yield a better value of f_{min} on the *test* set compared to the indices used for optimizing f_{min} itself; for example, see the final column corresponding to $4\times$ or $32\times$ compression.

C. MRI

Our final example considers a classical MRI application in medical imaging. For our illustrations, we use a data set consisting of 3-dimensional volumes of knee images, fully scanned on a GE clinical 3T scanner.[2] We focus our attention on a commonly-considered sampling technique based on subsampling the k-space in the x and y directions, while fully sampling the z direction. Note that we do not consider certain physical constraints such as those addressed in [7]; our focus is on using the data to illustrate our learning-based approach, rather than demonstrating direct real-world applicability.

We pick the first half of the patients in the given data set for training and test our results on the remaining 10 patients. For the nonlinear decoder, we use basis pursuit (BP) combined with complex dual-tree wavelets, since this has been shown to give superior performance compared to other basis/solver

combinations such as those involving total variation (TV) minimization [31]. Also following [31], in this work we are not concerned with the denoising aspect, but instead we are only comparing our reconstructions to a fully sampled noisy image which constitutes a ground truth that is used to compute the error and PSNR. To solve the BP algorithm, we made use of NESTA [32].

As a baseline for the subsampling map, we use the variable-density functions proposed in [3] and [4], which are determined by various parameters. Specifically, the former has a radius r within which full sampling is done and a polynomial degree d, and the latter has a radius r within which full sampling is done, a number of levels N, and two exponential parameters a and b.[3] We note that letting N be large and r be small in [4], we recover a very similar sampling pattern to that proposed in [11], which has only the parameters a and b.

For each of [3] and [4], we do a sweep over the parameters, generate 20 random subsampling patterns for each set of parameters, and finally choose the pattern with the best average PSNR on 100 fixed and randomly selected training images. For [3], we sweep over $r \in \{0, 0.025, \ldots, 0.5\}$ and $d \in \{0, 0.25, \ldots, 10\}$. For [4], we let the number of levels $N = 100$ be fixed and large, sweep over $r \in \{0, 0.025, \ldots, 0.5\}$ and $a \in \{0.05, 0.1, \ldots, 8, 9, \ldots, 25\}$, and compute b via a binary search in order to ensure that the total number of samples is exactly n. We note that the parameter sweeps for the nonlinear decoder are highly time consuming, taking considerably longer than our combinatorial optimization problems for the linear decoder.

Figure 7 illustrates the best-performing randomized indices vs. our learned set of indices in the k-space along the x and y directions. When optimized for the linear decoder, the indices of [3], [4] concentrate on low frequencies. While our strategies based on optimizing f_{avg} and f_{gen} (again using $g(\alpha) = 1 - (1 - \alpha)^2$) also do this to some extent, there is a stark contrast in the shape, since we do not restrict ourselves to patterns exhibiting circular symmetry.

Table III illustrates the overall test performance of each approach, in addition to the error obtained by the best adaptive (i.e., image dependent) n-sample approximations with respect to the k-space basis. Based on these numbers, the learning-based approach slightly outperforms the randomized variable-density based approach of Roman *et al.* [4], which in turn slightly outperforms that of Lustig *et al.* [3]. The best PSNR in each case is achieved by the indices corresponding to f_{avg}; with this choice, even the linear decoder leads to an improvement over [3] and [4] used with BP, while using our indices alongside BP provides a further improvement. Finally, based on Figure 8, it appears that the improvement of our indices is actually more significant in the relevant parts of the image where the knee is observed, with finer details being seen at 6.25% and 12.5% sampling rates.

V. CONCLUSION AND DISCUSSION

We have provided a principled learning-based approach to subsampling an orthonormal basis for structured signal

---

[2] Available at http://mridata.org/fullysampled

[3] We use slightly different notation to [4] to avoid clashes with our notation.

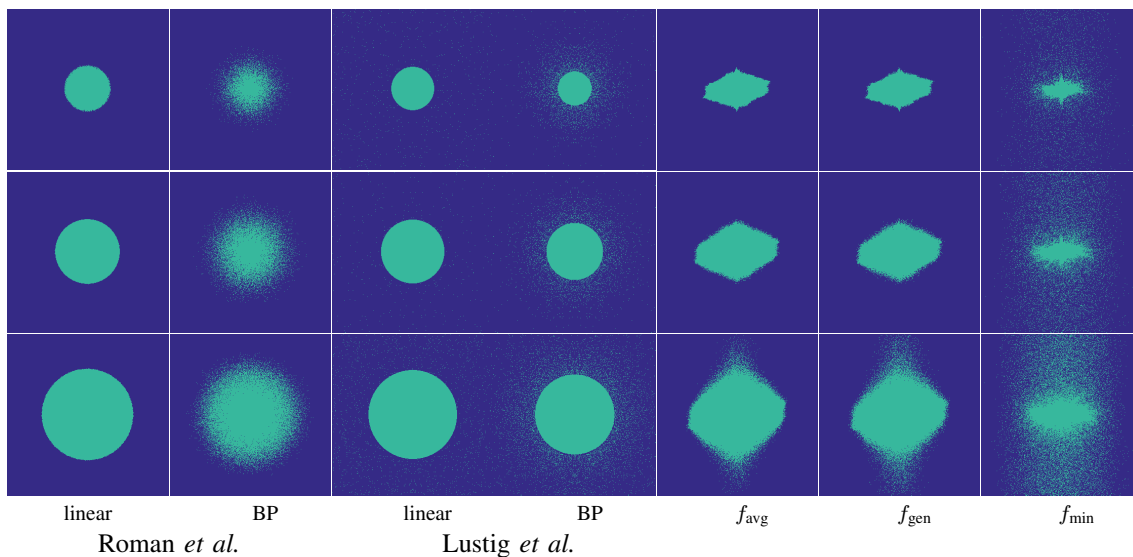|						
linear	BP	linear	BP	f_{avg}	f_{gen}	f_{min}
Roman *et al.*		Lustig *et al.*				

Fig. 7. The index sets of the tuned random variable sampling schemes [3], [4] vs. our learned-based approach with various optimization criteria. The sampling ratios of the three rows are 6.25%, 12.5% and 25%.

TABLE III
MRI: AVERAGE ℓ_2-ERRORS AND AVERAGE PSNR

Indices	Sampling rate					
	6.25%		12.50%		25%	
	ℓ_2	PSNR	ℓ_2	PSNR	ℓ_2	PSNR
Adaptive	0.371	25.299 dB	0.328	26.372 dB	0.261	28.362 dB
f_{avg} linear	0.399	24.673 dB	0.377	25.193 dB	0.339	26.126 dB
f_{gen} linear	0.404	24.549 dB	0.386	24.959 dB	0.352	25.757 dB
f_{min} linear	0.408	24.462 dB	0.385	24.969 dB	0.345	25.950 dB
Lustig et al. linear	0.404	24.561 dB	0.378	25.148 dB	0.340	26.095 dB
Roman et al. linear	0.401	24.625 dB	0.378	25.161 dB	0.339	26.123 dB
f_{avg} BP	0.398	24.699 dB	0.376	25.210 dB	0.338	26.145 dB
f_{gen} BP	0.399	24.689 dB	0.381	25.094 dB	0.347	25.901 dB
f_{min} BP	0.402	24.621 dB	0.380	24.105 dB	0.342	26.040 dB
Lustig et al. BP	0.401	24.640 dB	0.378	25.167 dB	0.341	26.076 dB
Roman et al. BP	0.405	24.526 dB	0.381	25.080 dB	0.340	26.085 dB

recovery, thus providing a powerful alternative to traditional approaches based on parametrized random sampling distributions. We proposed combinatorial optimization problems based on the average-case and worst-case energy captured in the training signals, obtained solutions via the identification of modularity and submodularity structures, and provided both deterministic and statistical generalization bounds. Our numerical findings reveal that our approach performs well on a variety of data sets, while having the desirable features of avoiding any reliance on randomization, and permitting the use of an efficient and scalable linear decoder.

While our numerical results support the use of the simple linear decoder in a variety of settings, we note that there are cases where it is unsuitable [8], [10]. In particular, to perform well, we need to capture most of the energy in $\mathbf{x}$ using only n measurements, and hence we require a degree of *compressibility in the measurement basis* $\mathbf{\Psi}$.

A simple example where the linear decoder is not suitable is when $\mathbf{x}$ is sparse in the canonical basis, whereas the measurements are taken in the Fourier basis. In this case, a large number of measurements is required to capture most of the energy, whereas for non-linear decoders, several results on the exact recovery of s-sparse signals are known using only $n = O(s \log^\nu p)$ measurements for some $\nu > 0$ [2]. See [10] for a related example with radial Fourier sampling, where

for a particular set of measurements it is observed that non-linear decoding is highly preferable. Based on these examples, extending our learning-based approach to non-linear decoders is a very interesting direction for further research.

A variety of other extensions would also be valuable, including the consideration of measurement noise, imperfectly-known training signals, more general recovery criteria beyond the ℓ_2-error, and further measurement constraints motivated by applications (e.g., see [7]).

ACKNOWLEDGMENT

REFERENCES

[1] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Sig. Proc. Magazine*, vol. 25, no. 2, pp. 21–30, March 2008.

[2] S. Foucart and H. Rauhut, *A mathematical introduction to compressive sensing*. Springer, 2013.

[3] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magn. Reson. Med.*, vol. 58, no. 6, pp. 1182–1195, 2007.

[4] B. Roman, A. Hansen, and B. Adcock, "On asymptotic structure in compressed sensing," *arXiv preprint arXiv:1406.4178*, 2014.

[5] W. B. Pennebaker and J. L. Mitchell, *JPEG: Still Image Data Compression Standard*. Springer US, 1993.

[6] B. Adcock, A. C. Hansen, C. Poon, and B. Roman, "Breaking the coherence barrier: A new theory for compressed sensing," *arXiv preprint arXiv:1302.0561v3*, 2013.

[7] C. Boyer, P. Weiss, and J. Bigot, "An algorithm for variable density sampling with block-constrained acquisition," *SIAM J. Imag. Sci.*, vol. 7, no. 2, pp. 1080–1107, 2014.

[8] E. J. Candes, "Modern statistical estimation via oracle inequalities," *Acta Numer.*, vol. 15, pp. 257–325, 2006.

[9] M. E. Halabi and V. Cevher, "A totally unimodular view of structured sparsity," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2015.

[10] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[11] Z. Wang and G. R. Arce, "Variable density compressed image sampling," *IEEE Trans. Image Proc.*, vol. 19, no. 1, pp. 264–270, 2010.

[12] G. Puy, P. Vandergheynst, and Y. Wiaux, "On variable density compressive sampling," *IEEE Sig. Proc. Letters*, vol. 18, no. 10, pp. 595–598, 2011.

[13] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Sig. Proc.*, vol. 55, no. 12, pp. 5695–5702, 2007.

[14] C. Hegde, A. Sankaranarayanan, W. Yin, and R. Baraniuk, "NuMax: A convex approach for learning near-isometric linear embeddings," *IEEE Trans. Sig. Proc.*, vol. PP, no. 99, pp. 1–1, 2015.

[15] A. Sadeghian, B. Bah, and V. Cevher, "Energy-aware adaptive bi-Lipschitz embeddings," in *Int. Conf. Samp. Theory and Apps.*, Bremen, Germany, 2013.

[16] B. Bah, S. Becker, V. Cevher, and B. Gözcü, "Metric learning with rank and sparsity constraints," in *IEEE Int. Conf. Acoustics, Speech, and Sig. Proc. (ICASSP)*, May 2014, pp. 21–25.

[17] A. Schrijver, *Combinatorial optimization: Polyhedra and efficiency*. Springer Science & Business Media, 2003, vol. 24.

[18] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[19] A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta, "Robust submodular observation selection," *J. Mach. Learn. Res.*, pp. 2761–2801, Dec. 2008.

[20] L. Baldassarre, N. Bhan, V. Cevher, A. Kyrillidis, and S. Satpathi, "Group-sparse model selection: Hardness and relaxations," 2013, http://arxiv.org/abs/1303.3207.

[21] A. Krause and D. Golovin, "Submodular function maximization," *Tractability: Practical Approaches to Hard Problems*, vol. 3, 2012.

[22] L. A. Wolsey and G. L. Nemhauser, *Integer and combinatorial optimization*. John Wiley & Sons, 2014.

[23] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM J. Comp.*, vol. 40, no. 6, pp. 1740–1766, 2011.

[24] Y. Filmus and J. Ward, "A tight combinatorial algorithm for submodular maximization subject to a matroid constraint," in *Ann. Symp. Found. Comp. Sci. (FOCS)*, 2012.

[25] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[26] Q. Tran-Dinh and V. Cevher, "A primal-dual algorithmic framework for constrained convex minimization," 2014, http://arxiv.org/abs/1406.5403.

[27] J. N. Laska, S. Kirolos, M. F. Duarte, T. S. Ragheb, R. G. Baraniuk, and Y. Massoud, "Theory and implementation of an analog-to-information converter using random demodulation," in *IEEE Int. Symp. Circ. Sys.*, 2007.

[28] F. Chen, A. P. Chandrakasan, and V. M. Stojanovic, "Design and analysis of a hardware-efficient compressed sensing architecture for data compression in wireless sensors," *IEEE J. Solid-State Circuits*, vol. 47, no. 3, pp. 744–756, 2012.

[29] M. Shoaran, M. H. Kamal, C. Pollo, P. Vandergheynst, and A. Schmid, "Compact low-power cortical recording architecture for compressive multichannel data acquisition," *IEEE Trans. Biomed. Circ. Sys.*, vol. 8, no. 6, pp. 857–870, Dec. 2014.

[30] L. Baldassarre, C. Aprile, M. Shoaran, Y. Leblebici, and V. Cevher, "Structured sampling and recovery of iEEG signals," in *IEEE Int. Workshop Comp. Adv. Multi-Sensor Adapt. Proc.*, 2015.

[31] A. Majumdar and R. K. Ward, "On the choice of compressed sensing priors and sparsifying transforms for MR image reconstruction: An experimental study," *Signal Processing: Image Communication*, vol. 27, no. 9, pp. 1035–1048, 2012.

[32] S. Becker, J. Bobin, and E. J. Candès, "NESTA: a fast and accurate first-order method for sparse recovery," *SIAM J. Imag. Sci.*, vol. 4, no. 1, pp. 1–39, 2011.
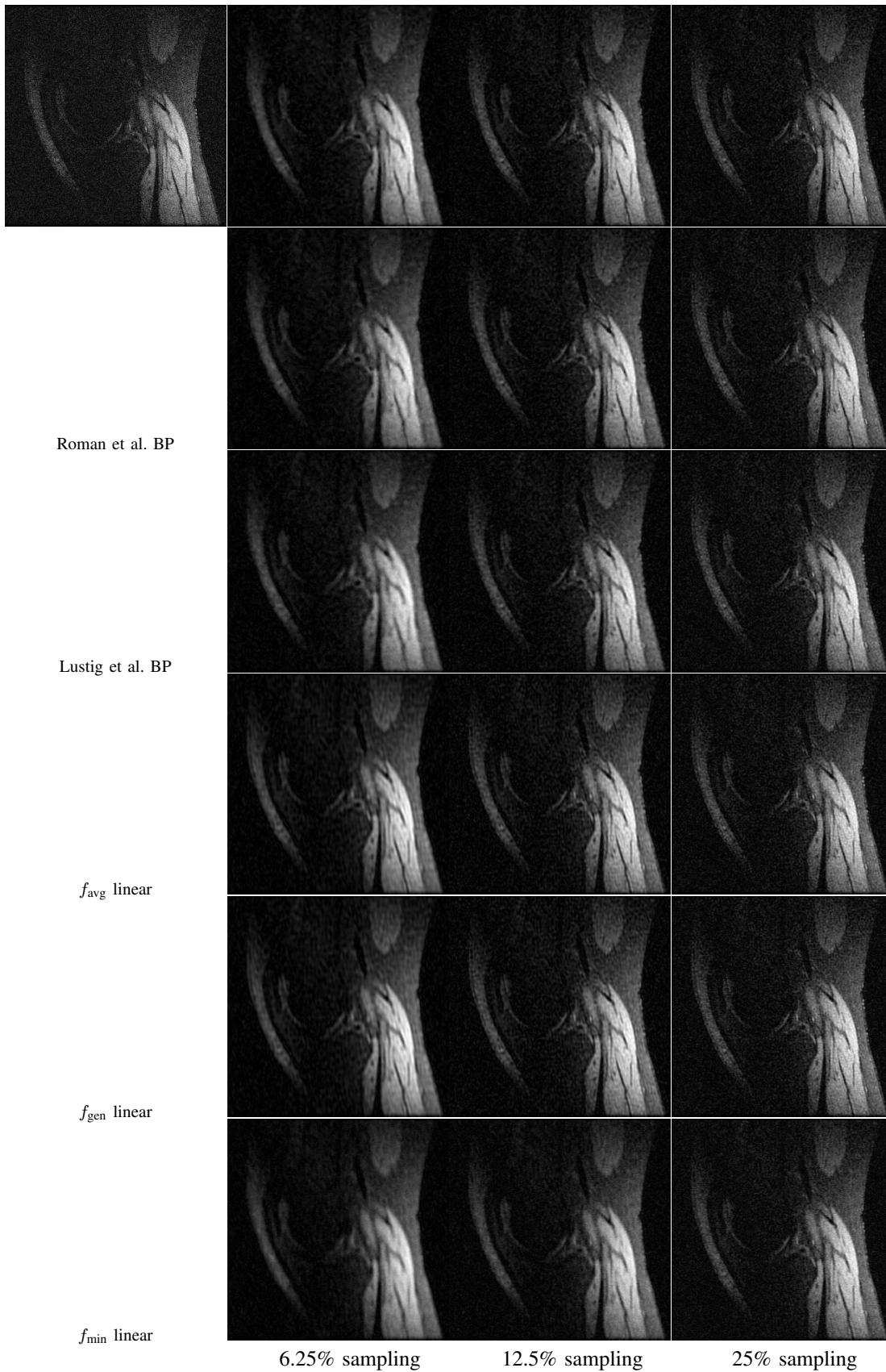
Fig. 8. MRI reconstruction example for the test patient 11. The top-left image corresponds to the original (fully sampled) data, and the remainder of the first row corresponds to using the best k coefficients in an image-adaptive fashion.