Constrained adaptive sensing

Mark A. Davenport, *Senior Member, IEEE,* Andrew K. Massimino, *Student Member, IEEE,*
Deanna Needell, *Member, IEEE,* and Tina Woolf

Abstract—Suppose that we wish to estimate a vector $\mathbf{x} \in \mathbb{C}^n$ from a small number of noisy linear measurements of the form $\mathbf{y} = \mathbf{Ax} + \mathbf{z}$, where $\mathbf{z}$ represents measurement noise. When the vector $\mathbf{x}$ is sparse, meaning that it has only s nonzeros with $s \ll n$, one can obtain a significantly more accurate estimate of $\mathbf{x}$ by adaptively selecting the rows of $\mathbf{A}$ based on the previous measurements provided that the signal-to-noise ratio (SNR) is sufficiently large. In this paper we consider the case where we wish to realize the potential of adaptivity but where the rows of $\mathbf{A}$ are subject to physical constraints. In particular, we examine the case where the rows of $\mathbf{A}$ are constrained to belong to a finite set of allowable measurement vectors. We demonstrate both the limitations and advantages of adaptive sensing in this constrained setting. We prove that for certain measurement ensembles, the benefits offered by adaptive designs fall far short of the improvements that are possible in the unconstrained adaptive setting. On the other hand, we also provide both theoretical and empirical evidence that in some scenarios adaptivity does still result in substantial improvements even in the constrained setting. To illustrate these potential gains, we propose practical algorithms for constrained adaptive sensing by exploiting connections to the theory of optimal experimental design and show that these algorithms exhibit promising performance in some representative applications.

Index Terms—sparsity, compressive sensing, adaptive sensing, constrained sensing, information bounds, optimal experimental design

I. INTRODUCTION

Suppose that we wish to estimate a sparse vector from a small number of noisy linear measurements. In the setting where the measurements are selected in advance (independently of the signal) we now have a rich understanding of both practical algorithms and the theoretical limits on the performance of these algorithms. A typical result from this literature states that for a suitable measurement design, one can estimate a sparse vector with an accuracy that matches the minimax lower bound up to a constant factor [1]. Such results have had a tremendous impact in a variety of practical settings.

M. A. Davenport and A. K. Massimino are with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332 USA (e-mail: mdav@gatech.edu, massimino@gatech.edu).

D. Needell is with the Department of Mathematical Sciences, Claremont McKenna College, Claremont, CA, 91711 USA (e-mail: dneedell@cmc.edu).

T. Woolf is with the Institute of Mathematical Sciences, Claremont Graduate University, Claremont, CA, 91711 USA (e-mail: tina.woolf@cgu.edu).

In particular, they provide the mathematical foundation for "compressive sensing," a paradigm for efficient sampling that has inspired a range of new sensor designs over the last decade.

A distinguishing feature of the standard compressive sensing paradigm is that the measurements are *nonadaptive*, meaning that a fixed set of measurements are designed and acquired without allowing for any possibility of adapting as the measurements begin to reveal the structure of the signal. While this can be attractive in the sense that it enables simpler hardware design, in the context of sparse estimation this also leads to some clear drawbacks. In particular, this would mean that even once the acquired measurements show us that portions of the signal are very likely to be zero, we may still expend significant effort in "measuring" these zeros! In such a case, by *adaptively* choosing the measurements, dramatic improvements may be possible.

Inspired by this potential, recent investigations have shown that we can often acquire a sparse (or compressible) signal via far fewer measurements or far more accurately if we choose them adaptively (e.g., see [2]–[5]). This body of work, which will be discussed in greater detail in Section I-B, demonstrates that adaptive sensing indeed offers the potential for dramatic improvements over nonadaptive sensing in many settings. However, the existing approaches to adaptive sensing, which rely on being able to acquire *arbitrary* linear measurements, cannot be applied in most real-world applications where the measurements must respect certain physical *constraints*. In this paper, our focus is on *constrained adaptive sensing*, where our measurements are restricted to be chosen from a particular set of allowable measurements. We will see that new algorithms are required and explore the theoretical limits within this more restrictive setting. Before describing the constrained adaptive setting in more detail, we first provide a brief review of existing approaches to nonadaptive and adaptive sensing of sparse signals.

A. Nonadaptive sensing

In the standard nonadaptive compressive sensing framework [6]–[9], we acquire a signal $\mathbf{x}$ via the linear measurements $\mathbf{y} = \mathbf{Ax} + \mathbf{z}$, where $\mathbf{A}$ is an $m \times n$ matrix representing the sensing system and $\mathbf{z}$ represents measurement noise. The goal is to design $\mathbf{A}$ so that m is smaller than n by exploiting the fact that $\mathbf{x}$ is *sparse* (or nearly sparse). Given a basis $\boldsymbol{\Psi}$, we say that a signal $\mathbf{x} \in \mathbb{C}^n$ is s-sparse if it can be represented by a linear combination of just s elements from $\boldsymbol{\Psi}$, i.e., we can write $\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\alpha}$, with $\|\boldsymbol{\alpha}\|_0 \leq s$, where $\|\boldsymbol{\alpha}\|_0 := |\mathrm{supp}(\boldsymbol{\alpha})|$ denotes the number of nonzeros in $\boldsymbol{\alpha}$. We will typically be interested in the case where $s \ll n$.

There is now a rich literature that describes a wide range of techniques for designing an appropriate $\mathbf{A}$ and efficient

algorithms for recovering $\mathbf{x}$. In much of this literature, the matrix $\mathbf{A}$ is chosen via randomized constructions that are known to satisfy certain desirable properties such as the so-called *restricted isometry property* (RIP).[1] Under the assumption that $\mathbf{A}$ satisfies the RIP (or that $\mathbf{A}\Psi$ satisfies the RIP in the case where $\Psi \neq \mathbf{I}$), if each entry of $\mathbf{z}$ is independent white Gaussian noise with variance σ^2 then one can show that techniques based on ℓ_1-minimization produce an approximation $\widehat{\mathbf{x}}$ satisfying

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq C \frac{n \log n}{\|\mathbf{A}\|_F^2} s\sigma^2, \tag{1}$$

where $C > 1$ is a fixed constant (e.g., see [9, pp. 35]). Note that this bound holds for *any* $\mathbf{x}$, and hence any SNR (even the worst-case). It is possible to obtain improved bounds that eliminate the $\log n$ factor when one assumes that the SNR is sufficiently large to ensure that the support is exactly recovered.

One can show that this result is essentially optimal in the sense that there is no alternative method to choose $\mathbf{A}$ or perform the reconstruction that can do better than this (up to the precise value of the constant C) [1]. In the event that the signal $\mathbf{x}$ is not exactly s-sparse, it is also possible to extend these results by introducing an additional term in the error bound that measures the error incurred by approximating $\mathbf{x}$ as s-sparse. See [9] and references therein for further details.

B. Adaptive sensing

A defining feature of the approach described above is that it is completely nonadaptive. When we consider the effect of noise, this nonadaptive approach might draw some severe skepticism. To see why, note that in the nonadaptive scenario, most of the "sensing energy" is used to measure the signal at locations where there is no information, i.e., where the signal vanishes. Specifically, one consequence of using the randomized constructions for $\mathbf{A}$ typically considered in the literature, or alternatively, any matrix satisfying the RIP, is that the available sensing energy (i.e., $\|\mathbf{A}\|_F^2$) is evenly distributed across all possible indices. This is natural since *a priori* we do not know where the nonzeros may lie, however, since most of the coordinates $\mathbf{x}_j$ are zero, it also means that the vast majority of the sensing energy is seemingly wasted. In other words, by design, the sensing vectors are approximately orthogonal to the signal, yielding a poor signal-to-noise ratio (SNR).

The idea behind adaptive sensing is that we should focus our sensing energy on locations where the signal is nonzero in order to increase the SNR, or equivalently, not waste sensing energy. In other words, one should try to learn as much as possible about the signal while acquiring it in order to design more effective subsequent measurements. Roughly speaking, one would like to (i) detect those entries which are nonzero or significant, (ii) progressively concentrate the

sensing vectors on those entries, and (iii) estimate the signal from such localized linear functionals. Such a strategy is employed by the *compressive binary search* and *compressive adaptive sense and search* strategies of [2] and [5]. These algorithms operate by examining successively smaller pieces of the signal to accurately determine the locations of signal energy. These techniques can yield dramatic improvements in recovery accuracy.

To quantify the potential benefits of an adaptive scheme, suppose that we observe

$$y_i = \langle \mathbf{a}_i, \mathbf{x} \rangle + z_i \tag{2}$$

where the z_i are independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma^2)$ entries and the $\mathbf{a}_i$ are allowed to depend on the measurement history $((y_1, \mathbf{a}_1), \cdots, (y_{i-1}, \mathbf{a}_{i-1}))$, with the only constraint being that $\sum_i \|\mathbf{a}_i\|_2^2 = \|\mathbf{A}\|_F^2$ is fixed. Consider a simple procedure that uses half of the sensing energy in a nonadaptive way to identify the support of an s-sparse vector $\mathbf{x}$ and then adapts to use the remaining half of the sensing energy to estimate the values of the nonzeros. If such a scheme identifies the correct support, then it is easy to show that this procedure can yield an estimate satisfying

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{2s}{\|\mathbf{A}\|_F^2} s\sigma^2. \tag{3}$$

If we contrast this result to that in (1), which represents the best possible performance in the nonadaptive setting, we see that this simple adaptive scheme can potentially improve upon the nonadaptive scheme by a factor of roughly $(n/s) \log n$, which represents a *dramatic* improvement in the typical scenario where $s \ll n$. Of course, this is predicated on the assumption that the first stage of support identification succeeds, which is not always the case.

A fundamental question is thus: *in practice, how much lower can the mean squared error (MSE) be when we are allowed to sense the signal adaptively?* The answer is a subtle one. In [10], it is shown that there is a fixed constant $C > 0$ such that

$$\inf_{\widehat{\mathbf{x}}} \sup_{\|\mathbf{x}\|_0 \leq s} \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq C \frac{n}{\|\mathbf{A}\|_F^2} s\sigma^2. \tag{4}$$

In other words, for even the best possible adaptive scheme there are s-sparse vectors for which our recovery error is bounded below by (4). This lower bound improves upon the nonadaptive performance (1) by only a factor of $\log n$, coming far short of the improvement that (3) indicates might be possible. Similar results are also obtained in [11]. These results are established by considering vectors that are so difficult to estimate that it is impossible to obtain a reliable estimate of their support, and so adaptive algorithms offer limited room for improvement over nonadaptive ones.

The result (4) does not say that adaptive sensing *never* helps. In fact, in practice it *almost always* does help. For example, when some or most of the nonzero entries in $\mathbf{x}$ are only slightly larger than the worst-case amplitude identified in [10], we *can* detect them sufficiently reliably to enable the dramatic improvements predicted in (3). More concretely,

---

[1]See Section II for a more detailed discussion of the RIP and its implications in the context of adaptive sensing. Note that the RIP is typically stated to require $\|\mathbf{A}\mathbf{x}\|_2 \approx \|\mathbf{x}\|_2$ for all s-sparse $\mathbf{x}$, which implies a fixed scaling for the matrix $\mathbf{A}$ where $\|\mathbf{A}\|_F^2 \approx n$. To ease the comparison with results that arise in contexts with alternative scalings, in the result stated in (1) we make no assumption on the scaling of $\mathbf{A}$ and merely require $\|\mathbf{A}\mathbf{x}\|_2 \approx \beta\|\mathbf{x}\|_2$ for some $\beta > 0$.

provided that σ^2 is not too large[2] relative to the nonzero entries of $\mathbf{x}$, a well-designed adaptive scheme, where the $\mathbf{a}_i$ are chosen sequentially as in [2], [5], can achieve

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq C' \frac{s}{\|\mathbf{A}\|_F^2} s\sigma^2 \tag{5}$$

for a fixed constant C', which represents an enormous improvement when $s \ll n$, and demonstrates that the potential benefits suggested in (3) can be realized in certain regimes.

We briefly note that these results are somewhat reminiscent of classical results from the field of information based complexity [8], [12]–[14] as well as more recent results in active learning [15]. Although this literature considers different observation models (e.g., noise-free observations of non-sparse signals), the general theme is that adaptivity is beneficial only in certain regimes (e.g., see [3]). In another direction, we also note that several authors have previously suggested Bayesian approaches to adaptive sensing that are highly relevant to the problems we study in this paper, but which currently lack much in the way of theoretical justification or understanding [16]–[18].

C. Constrained sensing

Up to this point, we have discussed results in which we essentially have complete freedom to design both the adaptive and nonadaptive measurements in an optimal fashion (that is, up to a constraint on $\|\mathbf{A}\|_F$). However, there are many applications where such freedom does not exist, and there are significant constraints on the kind of measurements that we can actually acquire. Such constraints arise in various hardware devices inspired by compressive sensing. For example, the single-pixel camera [19] acquires samples of an image by computing inner products with binary patterns. In this application we could still utilize adaptive measurements, but they must be binary. In other applications, we may be restricted to obtaining point samples of the signal of interest. For example, in standard sampling systems we are restricted to individually measuring each signal coefficient over time or space. Finally, in tomography and magnetic resonance imaging (MRI), as well as other medical imaging settings, we cannot acquire inner products with arbitrary linear functionals—we are limited to Fourier measurements.

In all of these settings, the measurements are *constrained*; we still have the flexibility to design measurements adaptively, but we can only select measurements from a fixed ensemble of predetermined measurements. Thus, the constrained setting will typically preclude the use of any of the adaptive sensing algorithms referenced above, and a new approach is required. Specifically, if we let $\mathcal{M} \subset \mathbb{C}^n$ denote the set of candidate measurement vectors, then the constrained adaptive sensing problem becomes one of sequentially selecting the rows $\mathbf{a}_i$ of our sensing matrix from the set $\mathcal{M}$. In this work, we assume the multiplicity of a particular measurement from $\mathcal{M}$ is allowed to be greater than one; that is, repeated measurements

are permitted. For the methods discussed in this paper, we will restrict our attention to the case where $\mathcal{M}$ is a finite set. For a majority of our discussion and examples, we will focus on the setting where $\mathcal{M} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n\}$ consists of rows from the Discrete Fourier Transform (DFT) matrix. We stress, however, that we need not require $|\mathcal{M}| = n$ in general.

With the restriction that the measurements be chosen from the DFT ensemble, Figure 1 illustrates the large potential difference between a completely nonadaptive sensing scheme, where the measurements are selected uniformly at random, and an "oracle" adaptive sensing scheme which uses a priori knowledge of the true locations of the nonzeros in a signal to carefully adapt the choice of measurement vectors to minimize the expected recovery error using the strategy outlined in Section III. In both cases, the Compressive Sampling Matching Pursuit (CoSaMP) [20] algorithm is used for the signal recovery. The median[3] squared error over 200 trials is displayed against the signal dimension n. Here, the signal is chosen to have a sparse Haar wavelet decomposition that is supported on a tree. The choice of a tree-sparse signal is motivated by the observation that natural images typically have a structured sparsity pattern in a wavelet domain due to correlations between scales. In these simulations, the noise level $\sigma^2 = 10^{-4}$ is held constant while the nonzero coefficients scale as $\sqrt{n}$ so that the per-measurement SNR is fixed. The number of measurements taken is set to be $m = 0.6n$ (rounding when necessary). See Section IV-A for further details regarding these simulations.

It is well-known that the DFT and Haar wavelet transforms are not incoherent, which implies that $\mathbf{A}\mathbf{\Psi}$ should not satisfy the RIP; hence, we would not expect blind nonadaptive sensing to do well in this setting. However, Figure 1 does illustrate the large potential for improvement over nonadaptive sensing. In this case, the adaptive algorithm can potentially improve the recovery error over nonadaptive sensing by roughly a factor of n/s, which represents a substantial gain when $s \ll n$. While we will see below that there are also nonadaptive strategies to address the coherence of Fourier and Haar which somewhat reduce the gap between adaptive and nonadaptive sensing in this case, we believe that this clearly illustrates the potential for adaptive sensing, even in the constrained setting.

D. Organization

The remainder of the paper is organized as follows. In Section II, we show a simple lower bound on the adaptive performance of systems limited to DFT measurements. We then generalize this result to the larger class of measurements satisfying the RIP. In both cases, the signal is assumed to be sparse in the canonical basis. In Section III, we give a method for measurement selection based on optimal experimental design. In Section IV, we provide simulations in a more realistic setting and display numerical results when Fourier measurements are used and the signal is assumed to be sparse

---

[2]For example, the compressive binary search procedure proposed in [2] succeeds in finding the location of the smallest nonzero entry of amplitude μ with probability $1 - \delta$ when $\mu^2/\sigma^2 > 16n \log(\frac{1}{2\delta} + 1)/\|A\|_F^2$. The result for the procedure in [5] is similar.

[3]Note that the *median* and *mean* curves exhibit the same overall behavior; however, we display the median error across all trials rather than the mean error throughout because the median, being a more robust measure, resulted in smoother curves with clearer trends between the methods.
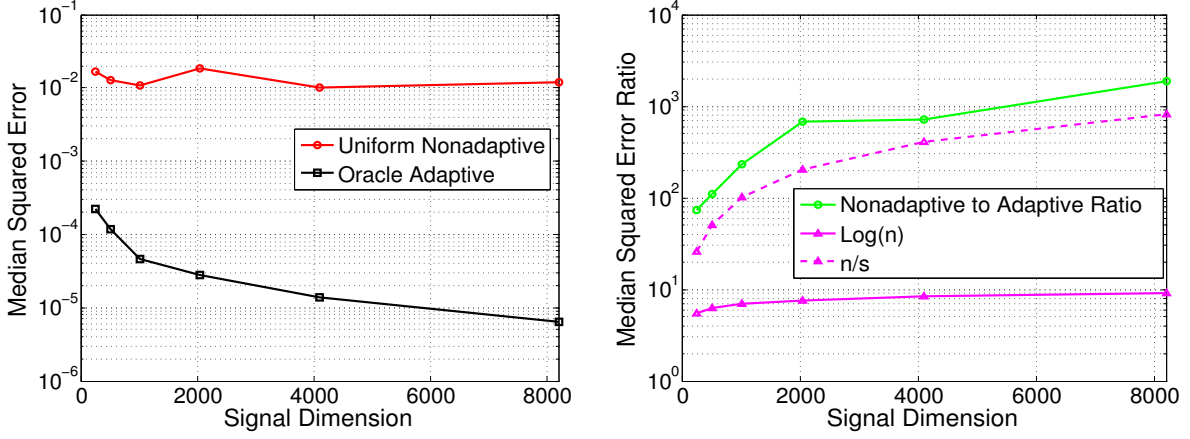
Fig. 1. (Left) The median squared error versus the signal dimension n for nonadaptive recovery with uniformly random selected measurements (red) and oracle adaptive recovery (black). (Right) The ratio (green) of the nonadaptive median squared recovery error to the oracle adaptive median squared recovery error versus the signal dimension n, with $\log n$ (solid magenta) and n/s (dashed magenta) included for reference.

in the Haar wavelet basis, for both synthetic and realistic signals. We also present some analytical justification using 1-sparse signals in this constrained adaptive setting. Finally, we conclude in Section V with a brief discussion.

II. LOWER BOUNDS ON ADAPTIVE PERFORMANCE

The main result of this section shows that adaptive sensing cannot offer substantial improvements over the nonadaptive scheme when the measurements are restricted to certain specific classes of ensembles and the signal is sparse in the canonical basis (i.e., $\mathbf{\Psi} = \mathbf{I}$). We first consider the Fourier ensemble, where the sensing vectors are chosen from the rows of the DFT matrix $\mathbf{F} \in \mathbb{C}^{n \times n}$, where $\mathbf{F}$ has entries

$$f_{jk} = \frac{1}{\sqrt{n}} \exp(-2\pi\sqrt{-1}jk/n) \tag{6}$$

for $j, k = 0, 1, \ldots, n-1$. In this constrained setting we have the following lower bound.

Theorem 1. *Under the adaptive measurement model of* (2), *where the* $\mathbf{a}_i$ *are chosen (potentially adaptively and allowing repeated measurements) by selecting rows from the DFT matrix* (6), *we have that*

$$\inf_{\widehat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \leq s \\ \|\mathbf{x}\|_2 \geq R}} \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq \frac{n}{m} s\sigma^2 \tag{7}$$

for any $R \geq 0$.

This shows that even using an optimal choice of sensing vectors, the recovery error is still proportional to $\frac{n}{m}s\sigma^2$, *even if we exclude the low-SNR setting* (by setting R to be large relative to σ). This is somewhat reminiscent of the main results of [10] and [11], which (in an unconstrained setting) establish minimax bounds of the form given in (4). However, a key difference is that in the unconstrained setting the worst-case error which defines the minimax rate is determined by the performance at a certain range of worst-case SNRs. Specifically, these bounds are obtained by constructing a "least favorable prior" where the nonzeros of $\mathbf{x}$ are near a specific

level,[4] and thus if we were to exclude these challenging $\mathbf{x}$ via the restriction that $\|x\|_2 \geq R$ as in (7), the bound in (4) would be dramatically lower – in particular, the gains shown in (5) could be realized [2], [5]. Thus, in a sense Theorem 1 is far more pessimistic than these results since it applies *no matter how large the SNR* – although given the incoherence of the DFT and the canonical bases, perhaps this is not that surprising. Finally, we note that for certain values of R it may be possible to obtain a slightly stronger version of Theorem 1 (by a $\log n/\log\log n$ factor) using the techniques in [21, Thm 6.1]. We do not pursue these refinements here.

Proof of Theorem 1. For any adaptive procedure $\widehat{\mathbf{x}}$, we let $\mathbf{F}'$ be the $m \times n$ sensing matrix consisting of the m adaptively chosen vectors from the rows of $\mathbf{F}$, and let $\mathbf{F}'_\Lambda$ denote the $m \times s$ submatrix of $\mathbf{F}'$ whose column indices correspond to the indices of the support Λ of $\mathbf{x}$. Using the rows of $\mathbf{F}'$ to acquire the measurements as in (2), we obtain $\mathbf{y} = \mathbf{F}'\mathbf{x} + \mathbf{z} = \mathbf{F}'_\Lambda \mathbf{x}_\Lambda + \mathbf{z}$. It is not difficult to show (e.g., see the Appendix of [1]) that

$$\inf_{\widehat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \leq s \\ \|\mathbf{x}\|_2 \geq R}} \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq \inf_{\widehat{\mathbf{x}}} \sup_{\substack{\mathbf{x}' \in \mathbb{R}^s \\ \|\mathbf{x}'\|_2 \geq R}} \mathbb{E}\|\widehat{\mathbf{x}}(\mathbf{F}'_\Lambda \mathbf{x}' + \mathbf{z}) - \mathbf{x}'\|_2^2,$$

where $\widehat{\mathbf{x}}(\cdot)$ takes values in $\mathbb{R}^s$.

To establish the bound in (7) we consider a sequence of least favorable prior distributions on $\mathbf{x}'$. The minimax risk is always larger than the Bayes risk under any prior, so this will establish a lower bound on the minimax risk. Towards this end, consider the prior on $\mathbf{x}'$ where $\mathbf{x}' \sim \mathcal{N}(0, \rho^2\mathbf{I})$, but where the distribution is truncated to be zero for $\|\mathbf{x}'\|_2 \leq R$ and re-scaled appropriately. Note that in the absence of this truncation, the Bayes risk would be given by

$$\sigma^2 \sum_{i=1}^{s} \left(\frac{\sigma_i(\mathbf{F}'_\Lambda)}{\sigma_i^2(\mathbf{F}'_\Lambda) + \frac{\sigma^2}{\rho^2}}\right)^2, \tag{8}$$

[4]This threshold is around $(\min_i x_i^2)/\sigma^2 \approx (n/m)\log s$.

where $\sigma_i(\mathbf{F}'_\Lambda)$ denotes the ith singular value of $\mathbf{F}'_\Lambda$. This follows from the fact that the Bayes estimator is given by

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\mathbf{F}'_\Lambda{}^T\mathbf{F}'_\Lambda + \tfrac{\sigma^2}{\rho^2}\mathbf{I})^{-1}\mathbf{F}'_\Lambda{}^T\mathbf{y}.$$

The result in (8) follows from the fact that for this estimator the expected squared error is given by

$$\sigma^2\|(\mathbf{F}'_\Lambda{}^T\mathbf{F}'_\Lambda + \tfrac{\sigma^2}{\rho^2}\mathbf{I})^{-1}\mathbf{F}'_\Lambda{}^T\|_F^2$$

which reduces to (8) via the application of standard properties of the singular value decomposition. We now note that for any $R \geq 0$, as $\rho^2 \to \infty$, the Bayes risk for the truncated prior will converge to that of (8), namely,

$$\sigma^2\|(\mathbf{F}'_\Lambda{}^T\mathbf{F}'_\Lambda)^{-1}\mathbf{F}'_\Lambda{}^T\|_F^2 = \sigma^2\sum_{i=1}^{s}\frac{1}{\sigma_i^2(\mathbf{F}'_\Lambda)}$$

Putting this all together, we have that

$$\inf_{\widehat{\mathbf{x}}}\sup_{\substack{\|\mathbf{x}\|_0\leq s\\\|\mathbf{x}\|_2\geq R}}\mathbb{E}\|\widehat{\mathbf{x}}-\mathbf{x}\|_2^2 \geq \sigma^2\sum_{i=1}^{s}\frac{1}{\sigma_i^2(\mathbf{F}'_\Lambda)}$$

$$\geq \sigma^2\frac{s^2}{\sum_{i=1}^{s}\sigma_i^2(\mathbf{F}'_\Lambda)}$$

$$= \sigma^2\frac{s^2}{\|\mathbf{F}'_\Lambda\|_F^2}$$

where the second inequality follows from Jensen's inequality. Since $\|\mathbf{F}'_\Lambda\|_F^2 = \frac{sm}{n}$, this completes the proof. $\square$

Our next result generalizes this type of lower bound to any ensemble whose submatrices satisfy the RIP with overwhelming probability. This statement is significant because it suggests that in some constrained situations, specifically many commonly studied in compressive sensing, there is little benefit from adaptivity. Formally, we define an RIP ensemble as follows.

Definition 1. *Let m be fixed. We say that an $n \times n$ matrix $\mathbf{A}$ with unit-norm rows is an* RIP *ensemble if for any $m' \geq m$ a random $m' \times n$ submatrix $\widetilde{\mathbf{A}}$, whose rows are uniformly chosen without replacement, satisfies*

$$0.5\frac{m'}{n}\|\mathbf{u}\|_2^2 \leq \|\widetilde{\mathbf{A}}\mathbf{u}\|_2^2 \leq 1.5\frac{m'}{n}\|\mathbf{u}\|_2^2, \qquad (9)$$

for all s-sparse $\mathbf{u}$ with probability $1 - \exp(-cn)$ (where c is such that $\exp(-cn) < 1/2n$).

Theorem 2 makes rigorous the claim that selecting rows intelligently from such a matrix yields no substantial improvement over a nonadaptive scheme.

Theorem 2. *Under the adaptive measurement model of (2), where the $\mathbf{a}_i$ are chosen (potentially adaptively and allowing repeated measurements) by selecting rows from an RIP ensemble as defined above, we have that*

$$\inf_{\widehat{\mathbf{x}}}\sup_{\substack{\|\mathbf{x}\|_0\leq s\\\|\mathbf{x}\|_2\geq R}}\mathbb{E}\|\widehat{\mathbf{x}}-\mathbf{x}\|_2^2 \geq \frac{sn}{3m^2}s\sigma^2 \qquad (10)$$

for any $R \geq 0$.

We note that one usually anticipates m to be on the order of $s\log n$, in which case this bound becomes

$$\mathbb{E}\|\widehat{\mathbf{x}}-\mathbf{x}\|_2^2 \geq \frac{sn}{3ms\log n}s\sigma^2 = \frac{n}{3m\log n}s\sigma^2,$$

which is roughly a factor of $\log^2 n$ lower than the upper bound in (1). This result shows that the recovery error with any adaptive measurements selected from some standard RIP ensemble again falls short of the possible gains shown in (3).

We also note here that the bound in Theorem 2 is worse by a factor of m/s than Theorem 1. However, we believe this is necessary due to the fact that the only assumption we place on $\mathbf{A}$ is that (9) holds with overwhelming probability; this is a much weaker requirement than insisting on DFT measurements as in Theorem 1. As a motivating example, fix some subset $\Lambda \subset \{1,\ldots,n\}$ of size s. Construct a matrix $\mathbf{A}$ by setting it to the DFT basis $\mathbf{F}$, with its first row modified in the following way: on Λ, multiply each entry by a factor of C where $C^2 = m/8s$ and off of Λ multiply each entry by a factor $c = \sqrt{(n-sC^2)/(n-s)}$. This yields a matrix $\mathbf{A}$ whose rows still have unit norm. In addition, one can show that for this new matrix $\mathbf{A}$, the property (9) still holds with the same probability for $\delta = 5/8$ for any $(m+1) \times n$ submatrix $\widetilde{\mathbf{A}}$. Construct an $(m+1)\times n$ matrix $\mathbf{A}'$ with the first row of $\mathbf{A}$ repeated $m-s+2$ times (since at least s rows need to be unique). Then one computes that $\|\mathbf{A}'_\Lambda\|_F^2 = \frac{s}{n}(s-1+\frac{m}{8s}(m-s+2)) \gtrsim m^2/n$. On the other hand, any matrix of the same size adaptively constructed from the DFT basis $\mathbf{F}$ has a squared Frobenius norm equal to $s(m+1)/n$. Thus we may indeed lose an m/s factor because of this weakened assumption.

Proof of Theorem 2. Let $\mathbf{A}'$ be the $m \times n$ matrix of the adaptively selected rows as in the theorem. Fix a support set Λ of size at most s. Let $\mathbf{A}'_\Lambda$ be the restriction of $\mathbf{A}'$ to the support set Λ. We will prove the result by showing a bound on the norm of the rows of $\mathbf{A}'_\Lambda$ which we obtain via an argument of contradiction. To that end, let $\mathbf{a}^\star$ be the row of $\mathbf{A}$ corresponding to the row of $\mathbf{A}'_\Lambda$ with the greatest Euclidean norm. Now consider drawing a random $(m+1)\times n$ submatrix $\widetilde{\mathbf{A}}$ of $\mathbf{A}$ that contains $\mathbf{a}^\star$ as a row. Then one can compute that any such submatrix $\widetilde{\mathbf{A}}$ satisfies (9) (with $m' = m+1$) with probability at least $1-\exp(-cn)n/(m+1) > 1-\exp(-cn)n$. Indeed, one sees formally that

$$\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (9)} \mid \mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})$$

$$= \frac{\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (9) and } \mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})}{\mathbb{P}(\mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})}$$

$$\leq \frac{\mathbb{P}(\widetilde{\mathbf{A}} \text{ does not satisfy (9))}}{\mathbb{P}(\mathbf{a}^\star \text{ is a row of } \widetilde{\mathbf{A}})}$$

$$\leq \frac{\exp(-cn)}{(m+1)/n}.$$

Now let $\widetilde{\mathbf{A}}^c$ be the remainder of the matrix, i.e., all rows of $\widetilde{\mathbf{A}}$ except row $\mathbf{a}^\star$. Similarly, one computes that any such matrix $\widetilde{\mathbf{A}}^c$ satisfies (9) (with $m' = m$) with probability at least $1 - \exp(-cn)n/(n-m) > 1 - \exp(-cn)n$. Thus *both* of these matrices satisfy (9) with probability at least $1-2\exp(-cn)n > 0$. For the sake of a contradiction, suppose

that $\|\mathbf{a}_\Lambda^\star\|_2^2 > 3m/n$. Observe that the signal $\mathbf{x} \in \mathbb{R}^n$ where $\mathbf{x}_\Lambda = \mathbf{a}_\Lambda^\star$ and padded with zeros off of the support Λ is an s-sparse signal. Then since both matrices satisfy (9), we must have that

$$\|\widetilde{\mathbf{A}}\mathbf{x}\|_2^2 = \|\widetilde{\mathbf{A}}^c\mathbf{x}\|_2^2 + |\langle\mathbf{a}_\Lambda^\star, \mathbf{x}\rangle|_2^2$$
$$\geq 0.5\frac{m}{n}\|\mathbf{x}\|_2^2 + \|\mathbf{x}\|_2^4$$
$$> \left(0.5\frac{m}{n} + \frac{3m}{n}\right)\|\mathbf{x}\|_2^2 \geq \frac{3.5m}{n}\|\mathbf{x}\|_2^2.$$

On the other hand, we must also have that

$$\|\widetilde{\mathbf{A}}\mathbf{x}\|_2^2 \leq 1.5\frac{m+1}{n}\|\mathbf{x}\|_2^2.$$

Combining these means that $\frac{3.5m}{n} \leq 1.5\frac{m+1}{n} \leq \frac{3m}{n}$, which is a contradiction. Thus, it must be that $\|\mathbf{a}_\Lambda^\star\|_2^2 \leq 3m/n$. Since $\mathbf{a}_\Lambda^\star$ is the largest row of $\mathbf{A}_\Lambda'$, we then have that

$$\|\mathbf{A}_\Lambda'\|_F^2 \leq m\|\mathbf{a}_\Lambda^\star\|_2^2 \leq \frac{3m^2}{n}.$$

Following the same argument as in the proof of Theorem 1, we thus have that

$$\inf_{\widehat{\mathbf{x}}} \sup_{\substack{\|\mathbf{x}\|_0 \leq s \\ \|\mathbf{x}\|_2 \geq R}} \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \geq \frac{s^2}{\|\mathbf{A}_\Lambda'\|_F^2}\sigma^2 \geq \frac{sn}{3m^2}s\sigma^2,$$

which completes the proof. □

III. ADAPTIVITY THROUGH OPTIMAL EXPERIMENTAL DESIGN

Although there are some settings where constrained adaptive sensing does not offer substantial improvement over the nonadaptive scheme, one can of course ask if there are other settings where notable gains are still possible. In order to address this question, we consider the simplified constrained adaptive sensing problem where we assume the support Λ of the signal $\mathbf{x}$ (with respect to the sparsity basis $\mathbf{\Psi}$) is known, or some estimate of the support is provided. How would we choose the measurements to best make use of this information, while still respecting that the measurements are constrained to be from the measurement ensemble $\mathcal{M}$?

Let $\{\mathbf{a}_i\}_{i=1}^m$ denote a sequence of length m with elements $\mathbf{a}_i \in \mathcal{M}$ corresponding to the measurements of $\mathcal{M}$ that are chosen.[5] Then, denote by $\mathbf{A}'$ the $m \times n$ matrix (recall $\mathcal{M} \subset \mathbb{C}^n$) whose ith row is $\mathbf{a}_i$. If $\Lambda = \mathrm{supp}(x)$, then it can be shown by following the arguments in the proof of Theorem 1 that the optimal MSE satisfies

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \|(\mathbf{A}'\mathbf{\Psi}_\Lambda)^\dagger\|_F^2\sigma^2$$
$$= \mathrm{tr}\left(((\mathbf{A}'\mathbf{\Psi}_\Lambda)^*\mathbf{A}'\mathbf{\Psi}_\Lambda)^{-1}\right)\sigma^2, \quad (11)$$

where $(\mathbf{A}'\mathbf{\Psi}_\Lambda)^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{A}'\mathbf{\Psi}_\Lambda$, σ^2 is the variance of the noise term as in (2), $\mathbf{\Psi}_\Lambda$ is the submatrix of $\mathbf{\Psi}$ restricted to the columns indexed by

[5] We use a *sequence* of elements from $\{1, \ldots, |\mathcal{M}|\}$ rather than a *subset* to emphasize that the m measurements from $\mathcal{M}$ need not be distinct. Note that in the general adaptive setting the *order* of the measurements is also important; however, in the context of this section there is only one batch of adaptive measurements and thus the order within this batch has no impact.

Λ, and $\mathbf{A}'\mathbf{\Psi}_\Lambda$ is assumed to have full (column) rank. Our goal is to find a length-m measurement sequence $\{\mathbf{a}_i\}_{i=1}^m$ that minimizes (11), which is equivalent to solving

$$\{\widehat{\mathbf{a}}_i\}_{i=1}^m = \underset{\{\{\mathbf{a}_i\}_{i=1}^m \mid \mathbf{a}_i \in \mathcal{M}\}}{\mathrm{argmin}} \mathrm{tr}\left(((\mathbf{A}'\mathbf{\Psi}_\Lambda)^*\mathbf{A}'\mathbf{\Psi}_\Lambda)^{-1}\right), \quad (12)$$

where $\mathbf{A}' = \mathbf{A}'(\{\mathbf{a}_i\}_{i=1}^m)$ is constructed as described above. Note that an essentially equivalent way to state (12) (up to a permutation of the measurements) is via the discrete optimization problem

$$\widehat{\mathbf{S}} = \underset{\substack{\text{diagonal matrices } \mathbf{S} \succeq 0 \\ s_{ii} \in \mathbb{Z}^+}}{\mathrm{argmin}} \mathrm{tr}\left(((\mathbf{A}\mathbf{\Psi}_\Lambda)^*\mathbf{S}\mathbf{A}\mathbf{\Psi}_\Lambda)^{-1}\right)$$
$$\text{subject to } \mathrm{tr}(\mathbf{S}) \leq m, \quad (13)$$

where $\mathbf{A}$ is the $|\mathcal{M}| \times n$ matrix containing all possible measurement vectors from $\mathcal{M}$ and $s_{ii} \in \mathbb{Z}^+$ forces each diagonal entry of $\mathbf{S}$ to be a non-negative integer (reflecting the multiplicity of each $\mathbf{a}_i$). Both (12) and (13) reflect the optimization problem that we would ideally like to solve. Unfortunately they are computationally demanding discrete optimization problems; hence, we instead consider the relaxation of (13)

$$\widehat{\mathbf{S}} = \underset{\text{diagonal matrices } \mathbf{S} \succeq 0}{\mathrm{argmin}} \mathrm{tr}\left(((\mathbf{A}\mathbf{\Psi}_\Lambda)^*\mathbf{S}\mathbf{A}\mathbf{\Psi}_\Lambda)^{-1}\right)$$
$$\text{subject to } \mathrm{tr}(\mathbf{S}) \leq m, \quad (14)$$

where the constraint $\mathrm{tr}(\mathbf{S}) \leq m$ ensures that the resulting "weighted" sensing matrix $\sqrt{\mathbf{S}}\mathbf{A}$ satisfies the "sensing energy" constraint $\|\sqrt{\mathbf{S}}\mathbf{A}\|_F^2 \leq m$ when the rows of $\mathbf{A}$ are normalized. Note that this is equivalent to the continuous design for the A-optimality criterion studied in the optimal experimental design literature [22].

Fortunately, (14) is a convex problem [23] and can be efficiently solved by a number of methods. Whereas the problem in (12) would tell us which measurements and how many of each to use from $\mathcal{M}$, (14) instead tells us, through the diagonal matrix $\widehat{\mathbf{S}}$ of weights, "how much" of each measurement to use. We simply weight each $\mathbf{a}_i$ by $\sqrt{\widehat{s}_{ii}}$, where $\widehat{s}_{ii}$ denotes the ith element on the diagonal of $\widehat{\mathbf{S}}$.

If, on the other hand, we use the measurement model where we must choose m unweighted measurements from $\mathcal{M}$, the practical use of $\widehat{\mathbf{S}}$ from (14) is less obvious. We experimented with several different (though likely sub-optimal) approaches to using the weights in $\widehat{\mathbf{S}}$, and the following method empirically seemed to produce the best results. In this work, we use a simple sampling scheme to obtain a discrete design. Specifically, we draw exactly m measurements, with replacement, according to the probability mass function

$$p_i = \frac{\widehat{s}_{ii}}{m}. \quad (15)$$

We guarantee that the resulting matrix $\mathbf{A}'$ is at least rank s by rejecting any construction for which this constraint is not satisfied. These m measurements then form the rows of the sensing matrix $\mathbf{A}'$.

IV. Case Study: Fourier Measurements of Wavelet Sparse Signals

The results of Section II demonstrate that adaptive sensing cannot offer substantial improvements over nonadaptive sensing for certain classes of measurement ensembles when the signal is sparse in the canonical basis. We next explore the case when $\boldsymbol{\Psi}$ is instead a wavelet basis and we acquire DFT measurements (this is indeed the setting of Figure 1, which suggests dramatic potential improvements from constrained adaptive sensing). This setting serves as a somewhat idealized model for a number of applications in tomography and other medical imaging since physical limitations would entail that we can only acquire DFT measurements, and realistic images are generally sparse with respect to wavelet bases [24]. In this setting we might receive one DFT measurement at a time, and from those, we can (potentially in real time) request the next DFT coefficient to be measured.

For our first two sets of experiments, we will assume the sparsity basis $\boldsymbol{\Psi}$ is the Haar wavelet basis. We will denote the $n \times n$ discrete Haar wavelet transform by $\mathbf{H}$, with entries h_{jk} for $j, k = 0, 1, \ldots, n-1$ and n is assumed to be some power of 2. When $j = 0$, we have

$$h_{0k} = \frac{1}{\sqrt{n}}. \tag{16}$$

For indices $j > 0$, we write $j = 2^p + q - 1$, where $p = \lfloor \log_2 j \rfloor$ and q are nonnegative integers, and define

$$h_{jk} = \frac{1}{\sqrt{n}} \begin{cases} 2^{p/2} & \frac{(q-1)n}{2^p} \le k < \frac{(q-0.5)n}{2^p} \\ -2^{p/2} & \frac{(q-0.5)n}{2^p} \le k < \frac{qn}{2^p} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Since, however, the Haar wavelet basis $\mathbf{H}$ is a sparsifying transformation, for a signal (or image) $\mathbf{x}$ we have that $\mathbf{Hx} = \boldsymbol{\alpha}$, with $\|\boldsymbol{\alpha}\|_0 \le s$. This means $\mathbf{x} = \mathbf{H}^* \boldsymbol{\alpha}$, where $\mathbf{H}^*$ denotes the adjoint of $\mathbf{H}$, for which $\mathbf{H}^* = \mathbf{H}^{-1}$ since $\mathbf{H}$ is unitary.

With this notation in hand and recalling that $\mathbf{F}$ is the $n \times n$ DFT, (11) becomes

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 \sigma^2, \tag{18}$$

where $\mathbf{F}'$ is the $m \times n$ sensing matrix consisting of the m adaptively chosen vectors from $\mathbf{F}$ and $\mathbf{H}_\Lambda^*$ is the $n \times s$ submatrix of $\mathbf{H}^*$ restricted to the columns indexed by $\Lambda = \mathrm{supp}(\boldsymbol{\alpha}) = \mathrm{supp}(\mathbf{Hx})$. Thus, we see that the optimal MSE depends on the correlations of the DFT and Haar basis elements. In a similar manner, in our last experiment, where the signal is an MRI image, we will assume the sparsity basis $\boldsymbol{\Psi}$ is the Daubechies wavelet with 3 vanishing moments (D6).

We now present a suite of numerical simulations in these settings that employ the relaxation (14) followed by the sampling scheme described in Section III to select a sequence of m DFT measurement vectors. We then follow with a short analysis for the simple case of 1-sparse signals.

A. Simulations

Here we present a practical implementation of adaptive sensing obtained via the relaxation (14) which we then compare with the results of traditional nonadaptive sensing. To implement (14), we use the Templates for First-Order Conic Solvers (TFOCS) software package [25], [26].

For our first two sets of experiments, we set $\mathcal{M}$ to be the ensemble of n measurements from the $n \times n$ DFT matrix $\mathbf{F}$. We define $\mathbf{x}$ to be a 10-sparse signal in the Haar wavelet basis (i.e., $\boldsymbol{\Psi} = \mathbf{H}^\star$) with the values on the support of $\boldsymbol{\alpha}$ distributed i.i.d as $\mathcal{N}(\sqrt{n}, 1)$ and the measurement noise $\mathbf{z}$ is distributed as i.i.d. $\mathcal{N}(0, 10^{-4})$.[6] Unless otherwise stated, the signal is of dimension $n = 1024$. We consider signals whose support is chosen uniformly at random, and also those whose support obeys a tree structure. Briefly, in the latter case the support is organized on a binary tree, plus an extra node at the top. The first scaling (or lowest frequency) coefficient has just one child; the second and further wavelet coefficients have two children each. This model is characteristic of natural images which tend to have inter-scale correlations (see [27], [28] for similar wavelet-tree constructions). An s-sparse support is filled by choosing the first scaling location, and then in each of the $s - 1$ remaining rounds, choosing one node randomly among the unfilled nodes which currently have a chosen parent.

Nonadaptive sensing. Due to the lack of incoherence between the DFT and Haar bases, it has been observed (and recently theoretically shown [29], [30]) that so-called *Variable-Density Sampling* (VDS) is often preferable to standard uniform random selection of DFT measurements. In VDS[7], sampling can be concentrated on the lower frequencies, producing superior recovery results. We test recovery using either ℓ_1-minimization [31]–[33] or the greedy pursuit CoSaMP [20].

Adaptive sensing. In the more realistic setting, we employ a simple strategy which uses $m/2$ nonadaptive measurements (using either VDS or uniform sampling) to construct an estimate of Λ. This is done by executing either ℓ_1-minimization (followed by thresholding) or CoSaMP. We then solve the relaxation (14) using this estimated support, and the remaining $m/2$ measurements are selected adaptively[8] using the distribution given by (15). To recover the signal, either ℓ_1-minimization or CoSaMP is again used to obtain an updated estimate $\hat{\Lambda}$ using all m measurements.[9] The final signal coefficient estimate is calculated as $\widehat{\boldsymbol{\alpha}}_{\hat{\Lambda}} = (\mathbf{F}'\mathbf{H}_{\hat{\Lambda}}^*)^\dagger \mathbf{y}$, where $\mathbf{y}$ is the m-dimensional vector of measurements, $\mathbf{F}'$ is the $m \times n$ vector of DFT measurements selected, and $\mathbf{H}_{\hat{\Lambda}}^*$ is the $n \times 10$ submatrix of $\mathbf{H}^*$ restricted to the columns indexed by $\hat{\Lambda}$. One could alternatively use the signal estimate returned directly from the recovery algorithm, which we have observed to perform similarly to (or only slightly worse than) our implemented method.

---

[6] We have found the adaptive procedure to be robust to the noise level, and compare similarly to the corresponding nonadaptive procedure even for larger noise levels.

[7] Following the experiments in [29], we also do not apply any preconditioning to the sensing matrix.

[8] Note that these $m/2$ *adaptively* selected measurements are only adapted to the first $m/2$ measurements, but are nonadaptive with respect to each other. That is, only one instance of adaptive measurement selection is being performed. Although in a different context, a similar two-stage approach is also taken in [15].

[9] Using dependent measurements is of course not justified theoretically, but we found unsurprisingly that using all m measurements gave better empirical results.
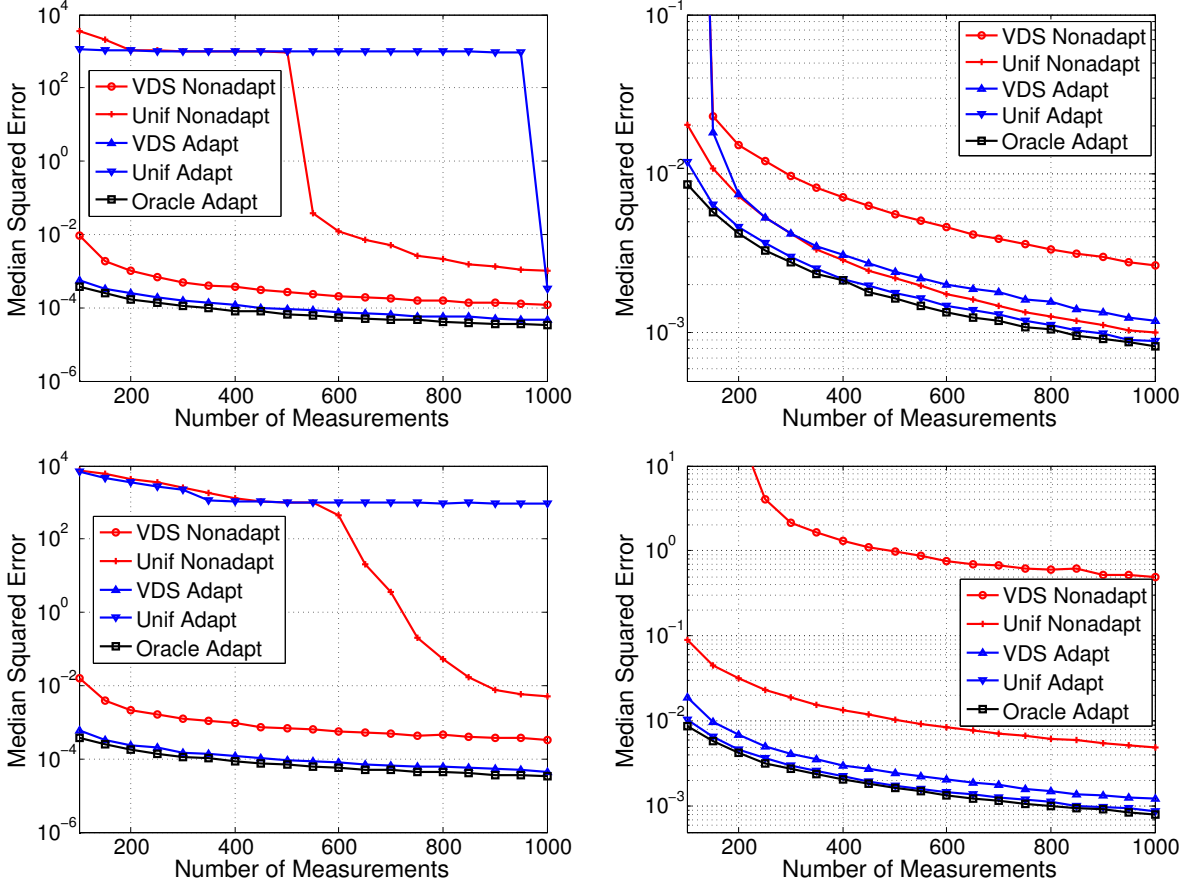
Fig. 2. (Large measurement regime) The median squared error versus the number of measurements m when the nonzero locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The nonadaptive (red) and adaptive (blue) recovery is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP (top) or ℓ_1-minimization (bottom) is used; the oracle adaptive (black) recovery is also included for comparison.

Oracle adaptive sensing. For sake of comparison, we also consider the case where the true support Λ of the signal is known a priori, and the measurements are selected as in the adaptive sensing case using this Λ. Recovery is then performed simply by applying the pseudoinverse: $\widehat{\boldsymbol{\alpha}}_\Lambda = (\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger \mathbf{y}$.

Figure 2 compares recovery results over 1000 trials for nonadaptive, adaptive, and oracle adaptive sensing versus the number of measurements m, where m ranges between 100 and 1000. We see that when the signal is supported on a tree, uniform sampling performs poorly for both nonadaptive and adaptive sensing, as might be expected. The performance of the uniform sampling methods can be understood via the empirical observation that in this case we require roughly 500 measurements before we can reliably estimate the support. When using the CoSaMP algorithm, the sudden improvement at $m \approx 500$ for uniform nonadaptive and at $m \approx 1000$ for uniform adaptive (which uses $m \approx 500$ measurements for support identification) corresponds to the threshold where more than half of the trials resulted in a correct support recovery. In contrast, sampling with VDS offers dramatic improvements for both nonadaptive and adaptive sensing with either reconstruction algorithm, with adaptive sensing performing almost as well as the oracle. In this case, VDS is already capturing much of the potential improvement offered

by adaptivity because the energy of the signal is heavily biased towards the lower frequencies, although adaptivity still results in somewhat improved performance. In contrast to the tree-sparse case, when the signal support is selected randomly, uniform nonadaptive sampling actually performs better than VDS, whereas adaptive sensing performs similarly regardless of the type of nonadaptive measurements taken. Thus if one is not sure of the signal structure in general, adaptive sensing can offer improvements in either case. This flexibility represents one of the main advantages of adaptive sensing.

Figure 3 studies the same setting as Figure 2 when using CoSaMP for recovery, but focuses on the small measurement regime. These results illustrate that there are regions, however narrow, where the nonadaptive method can succeed while the adaptive method fails. This is expected due to the nature of the adaptive scheme, where only $m/2$ measurements are utilized to identify a support estimate. At some point, the support can be sufficiently estimated with m, but not $m/2$, measurements. For tree-sparse signals, nonadaptive sensing with VDS measurements outperforms adaptive sensing with VDS nonadaptive measurements when $m \approx 50$. For uniformly sparse signals, we see this behavior even more clearly for both VDS and uniform sampling.

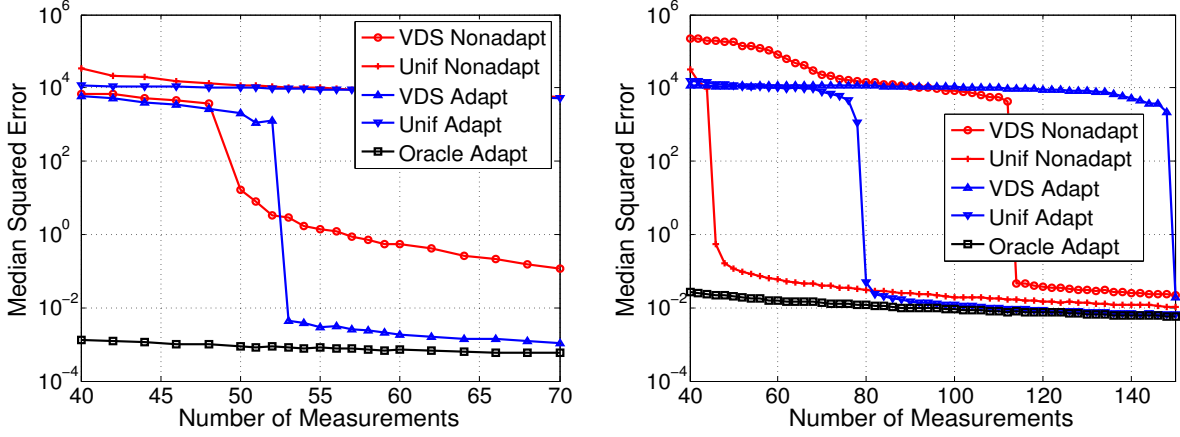The results of our second simulation are shown in Figure 4,

Fig. 3. (Small measurement regime) The median squared error versus the number of measurements m when the nonzero locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The nonadaptive (red) and adaptive (blue) recovery is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP is used; the oracle adaptive (black) recovery is also included for comparison.
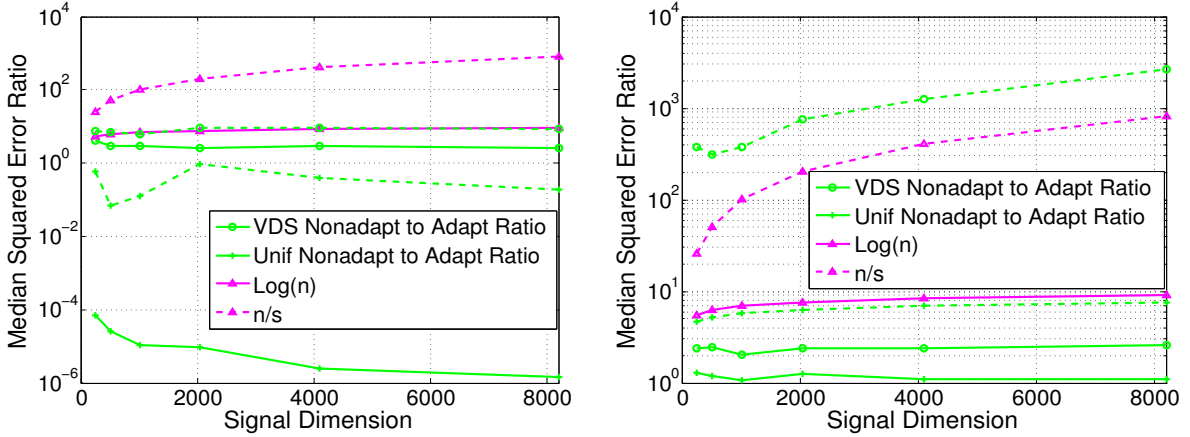

Fig. 4. The ratio (green) of the nonadaptive median squared recovery error to the adaptive median squared recovery error versus the signal dimension n when the support locations of $\boldsymbol{\alpha}$ are selected on a sparse tree (left) or uniformly at random (right). The ratio is shown when either VDS or uniform sampling is used for the nonadaptive measurements, and CoSaMP (solid line) or ℓ_1-minimization (dashed line) is used. The curves for $\log n$ (solid magenta) and n/s (dashed magenta) are included for comparison.

where we compare the ratio of nonadaptive to adaptive sensing recovery over 200 trials against the dimension n of the signal $\mathbf{x}$; the number of measurements used is always $m = 0.6n$ (rounding when necessary). We note that since the norms of $\boldsymbol{\alpha}$ and $\mathbf{z}$ both scale with n, the SNR remains roughly the same for all signal dimensions n. We observe similar results as Figure 2, demonstrating the behavior holds as a function of dimension.

In our last experiment, we evaluate our adaptive approach on real images. This scenario differs from previous experiments in two key aspects. First, the signal of interest is a two-dimensional (2D) image, not a one-dimensional vector, and thus we use 2D DFT measurements and a 2D discrete wavelet transform as the sparsity basis. Second, the image is not exactly sparse in any wavelet basis. Hence, when estimating the sparse support we introduce an additional (non-Gaussian) source of error, the contribution of the off-support wavelet coefficients. We note that the choice of the parameter s, which we have not attempted to optimize, can have an impact on

signal reconstruction.

The image we use, `brain.mat`[10], is rescaled to be 64×64, and is shown in Figure 5. We use the Daubechies wavelet with 3 vanishing moments (D6) in a full 2D decomposition (i.e., $\log_2 64 = 6$ levels). We set the parameter $s = 1000$, which we again note was not tuned nor optimized. Additionally, we introduce white Gaussian noise at the level of $\sigma = 0.01$ to each measurement.

The experiment proceeds as follows: in the nonadaptive case, m measurements are taken according to VDS. The set of recovered wavelet coefficients are obtained using ℓ_1-minimization and the image is reconstructed using the inverse wavelet transform. Note that the output of ℓ_1-minimization is not necessarily exactly s-sparse. The assumed sparsity s guides our choice of the ℓ_2 error term constraint, but we did no thresholding afterwards. We evaluate performance by the median peak signal-to-noise ratio (PSNR) in dB over 50 trials.

[10]Obtained from http://www.eecs.berkeley.edu/~mlustig/CS.html.

9

In the adaptive case, $m/2$ VDS measurements are taken as in the previous nonadaptive case. Then, via the ℓ_1-minimization reconstruction, we determine the estimated top s wavelet coefficients in each of 50 trials. We choose the trial with accuracy (in terms of the number of correctly identified top s wavelet coefficients) closest to the median accuracy. Utilizing the size s support estimate identified, we solve the relaxation (14) and select the remaining $m/2$ measurements adaptively. Finally, we recover the signal via ℓ_1-minimization, and, as before, reconstruct the final wavelet coefficients using the pseudoinverse. Again, we evaluate performance by the median PSNR over 50 trials of the adaptive measurement selection.

The adaptive and nonadaptive recovered images of the single trial with the closest to median PSNR performance using a total of $m = 3000$ measurements are given in Figure 5. Notice that the PSNR of the adaptive strategy is 28.02 dB, which exceeds the PSNR of 25.03 dB of the nonadaptive strategy. Visually, the adaptive strategy more closely resembles the original image. The median PSNR as the number of measurements is varied is shown in Figure 6. The plot shows that as the number of measurements reaches a certain level (roughly above 2000 measurements), the two-stage adaptive approach begins to exceed the method which is purely non-adaptive. Hence, as long as enough nonadaptive measurements are taken to obtain a sufficient support estimate, the adaptive procedure can improve image reconstruction quality.

We note that adaptive approaches to medical imaging have also been studied using an alternative Bayesian model for sampling optimization [34]–[36]. The work [36] studies the optimization of sequential sampling over stacks of neighboring image slices. In future work, it would be interesting to extend our proposed adaptive sampling scheme to this setting. Our method, however, is a framework for general adaptive sensing, not tuned specifically for medical imaging.

B. Analysis of the 1-sparse case

We now provide some analytical justification explaining why adaptive sensing can achieve a lower MSE than nonadaptive sensing for the Haar wavelet basis with DFT measurements, but show that the largest gains are realized for a small fraction of the possible signal support sets. We consider the simple case when $s = 1$ and the support is eventually known (either by oracle or by utilizing some method for estimation, as in the above experiments), and use this toy problem as motivational justification for the general setting. If we denote the s singular values of $\mathbf{F}'\mathbf{H}_\Lambda^*$ by $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s > 0$, then in general we want to minimize

$$\|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 = \sum_{i=1}^{s} \frac{1}{\sigma_i^2}. \tag{19}$$

When $s = 1$, (19) becomes $\|(\mathbf{F}'\mathbf{H}_\Lambda^*)^\dagger\|_F^2 = \frac{1}{\sigma_1^2}$. However, minimizing this quantity is the same as maximizing $\|\mathbf{F}'\mathbf{H}_\Lambda^*\|_F^2 = \sigma_1^2$. It is easy to see that $\|\mathbf{F}'\mathbf{H}_\Lambda^*\|_F^2$ is maximized when a measurement of $\mathbf{F}$ that is most correlated with $\mathbf{H}_\Lambda^*$ is chosen for every measurement in $\mathbf{F}'$. That is, if such a row can be identified, the best way to sense 1-sparse signals

Original

Nonadaptive,
$\mathrm{PSNR} = 25.03\,\mathrm{dB}$.

Adaptive,
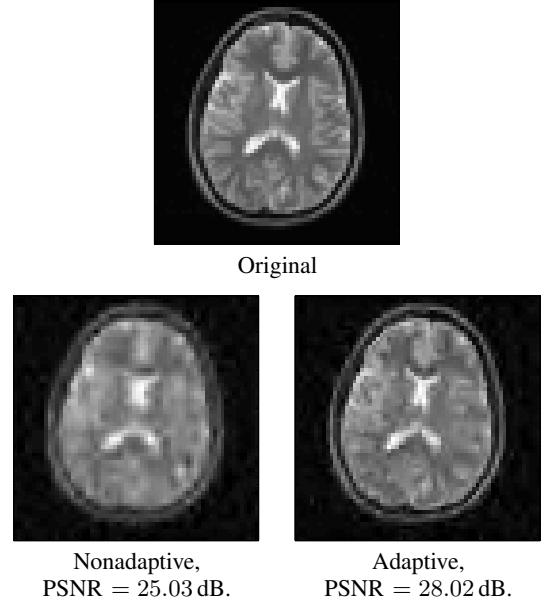$\mathrm{PSNR} = 28.02\,\mathrm{dB}$.

Fig. 5. (Top) The 64×64 `brain.mat` image used in our medical imaging experiments. Reconstructed images with the closest to median PSNR among the 50 trials of nonadaptive (bottom left) and adaptive (bottom right) sensing with $m = 3000$ measurements.
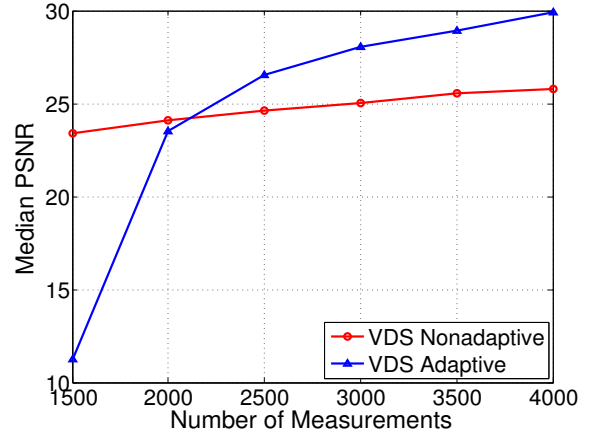

Fig. 6. The median PSNR versus the number of measurements m over 50 trials of nonadaptive and adaptive sensing of the 64×64 `brain.mat` image.

adaptively once the support is known is to simply repeat that measurement until the number of allotted measurements has been reached. Note that $\mathbf{F}'\mathbf{H}_\Lambda^*$ is $m \times 1$ in this case, and is still full rank when selecting the measurements in this way; thus, the theory leading to (18) still holds. In this setting, we can determine explicitly what the MSE looks like, and provide bounds on the MSE that depend on the support location of the 1-sparse signal. The result assuming a known support is provided in Theorem 3. The result in the more realistic context of adaptive sensing, where the first half of the measurements are selected nonadaptively, immediately follows and is provided in Corollary 1.

Theorem 3. *Denote by* $\mathbf{x} = \mathbf{H}^*\boldsymbol{\alpha}$ *the signal of interest, and suppose* $\mathbf{x}$ *becomes 1-sparse after applying the Haar wavelet transformation* $\mathbf{H}$ *(that is,* $\mathbf{H}\mathbf{x} = \boldsymbol{\alpha}$ *and* $\|\boldsymbol{\alpha}\|_0 = 1$*). Let* $\mathrm{supp}(\boldsymbol{\alpha}) = \Lambda$*, and suppose the support* Λ *is completely*

known. Suppose we measure repeatedly with a particular measurement from the $n \times n$ DFT $\mathbf{F}$ defined in (6); denote this measurement by $\mathbf{f}_j$, where $j \in \{0, 1, \ldots, n-1\}$ is some row index. Then, our observations are of the form

$$y_i = \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \boldsymbol{\alpha}_\Lambda \rangle + z_i, \qquad (20)$$

for $i = 1, \ldots, m$, where the noise z_i are i.i.d. $N(0, \sigma^2)$. Then the MSE is given by

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2} \sigma^2, \qquad (21)$$

and is bounded by

$$\frac{\sigma^2}{m} \leq \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{2m}, \qquad (22)$$

where the expectation is taken with respect to $\mathbf{z}$.

Note that in standard compressive sensing when we rely on the RIP, the DFT matrix $\mathbf{F}$ is normalized by $\frac{1}{\sqrt{m}}$ rather than $\frac{1}{\sqrt{n}}$. If we make this normalization, the bound in (22) becomes

$$\frac{\sigma^2}{n} \leq \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{\sigma^2}{2}.$$

Including pre-conditioning and other scalings of course yields an analogous bound.

The proof of Theorem 3 relies on three lemmas that provide bounds for the term $|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ appearing in the MSE in (21). Specifically, one term of interest is

$$\min_{\Lambda \in \{0, \ldots, n-1\}} \max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|. \qquad (23)$$

The maximization over j corresponds to selecting the best DFT measurement $\mathbf{f}_j$, and the minimization accounts for the worst case signal (i.e., the worst case support Λ). On the other hand, we also want to obtain a value that represents the best case signal so we are also interested in

$$\max_{\Lambda \in \{0, \ldots, n-1\}} \max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|. \qquad (24)$$

Before proving Theorem 3, let us set some notation. The Haar wavelet transform matrix $\mathbf{H}$, defined in (16) and (17) consists of *blocks* of consecutive rows with the same nonzero entry magnitudes. Let $1 \leq a \leq \log_2 n$ denote the block of $\mathbf{H}$, where $a = 1$ corresponds to the $\frac{n}{2}$ rows indexed by $j = \frac{n}{2}, \ldots, n-1$ (i.e., the "bottom" half of $\mathbf{H}$), $a = 2$ corresponds to the $\frac{n}{4}$ rows indexed by $j = \frac{n}{2} - \frac{n}{4}, \ldots, \frac{n}{2} - 1$, and so on. Similarly, for $\mathbf{H}^*$, instead of blocks of *rows*, we have blocks of *columns*; the block corresponding to $a = \log_2 n$ represents the lowest frequency wavelets, and the block corresponding to $a = 1$ represents the highest frequency wavelets.

Proof of Theorem 3. This proof requires the following three lemmas. Lemmas 1 and 2 are used to prove Lemma 3, and Lemma 3 is used to complete the proof of Theorem 3. The lemmas can be derived using elementary trigonometric bounds, and we omit the proofs here.

Lemma 1. *Fix $j \in \mathbb{Z}$ where $1 \leq j \leq n-1$ and let $a = 1, \ldots, \log_2 n$. Choose $k \in \mathbb{Z}$, $0 \leq k \leq n - \frac{2^a}{2}$. Then*

$$\left| \sum_{q=k}^{k + \frac{2^a}{2} - 1} e^{-2\pi i j q/n} \right| = \sqrt{\frac{1 - \cos(\frac{2^a \pi j}{n})}{1 - \cos(\frac{2\pi j}{n})}}. \qquad (25)$$

Lemma 2. *Let $\mathbf{f}_j$, $j \in \{0, \ldots, n-1\}$, be row j from the $n \times n$ DFT and let $\mathbf{H}_\Lambda^*$ be the inverse discrete Haar wavelet transform restricted to the column indexed by Λ. Let $a = 1, \ldots, \log_2 n$ denote the block of $\mathbf{H}^*$ and let $\Lambda \in \{1, 2, \ldots, n-1\}$, $|\Lambda| = 1$, be a column in the set corresponding to block a. Then,*

$$|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle| = \frac{1}{\sqrt{n2^{a-1}}} \frac{1 - \cos(\frac{2^a \pi j}{n})}{\sqrt{1 - \cos(\frac{2\pi j}{n})}}, \qquad (26)$$

where $j = 1, \ldots, n-1$. When $j = 0$,

$$|\langle \mathbf{f}_0, \mathbf{H}_\Lambda^* \rangle| = \begin{cases} 1 & \Lambda = 0 \\ 0 & \Lambda \in \{1, \ldots, n-1\}. \end{cases}$$

When $\Lambda = \{0\}$,

$$|\langle \mathbf{f}_j, \mathbf{H}_0^* \rangle| = \begin{cases} 1 & j = 0 \\ 0 & j = 1, 2, \ldots, n-1. \end{cases}$$

Lemma 3. *Let $\mathbf{f}_j$, $j \in \{0, \ldots, n-1\}$, be a row from the $n \times n$ DFT matrix $\mathbf{F}$ and let $\mathbf{H}_\Lambda^*$ be the inverse discrete Haar wavelet transform restricted to the column indexed by Λ. Let $\Lambda \in \{0, 1, \ldots, n-1\}$ so that $|\Lambda| = 1$. Then*

$$\min_{\Lambda \in \{0, \ldots, n-1\}} \max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle| = \sqrt{\frac{2}{n}} \qquad (27)$$

and

$$\max_{\Lambda \in \{0, \ldots, n-1\}} \max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle| = 1. \qquad (28)$$

Since $|\Lambda| = 1$, $\boldsymbol{\alpha}_\Lambda$ is just a scalar, so that the measurements (20) can be written as

$$y_i = \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \boldsymbol{\alpha}_\Lambda + z_i. \qquad (29)$$

This can be concisely written as

$$\mathbf{y} = \mathbf{A}\boldsymbol{\alpha}_\Lambda + \mathbf{z}, \qquad (30)$$

where $\mathbf{A}$ is the m-dimensional column vector with each entry equal to $\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle$. To estimate $\boldsymbol{\alpha}_\Lambda$, we apply $\mathbf{A}^\dagger$ to $\mathbf{y}$. In this case, $\mathbf{A}^\dagger$ is an m-dimensional row vector, with each entry equal to $\frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}$. Therefore,

$$\widehat{\boldsymbol{\alpha}}_\Lambda = \mathbf{A}^\dagger \mathbf{y} = \mathbf{A}^\dagger(\mathbf{A}\boldsymbol{\alpha}_\Lambda + \mathbf{z})$$

$$= \begin{bmatrix} \frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} & \cdots & \frac{1}{m\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} \end{bmatrix} \left(\begin{bmatrix} \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \\ \vdots \\ \langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle \end{bmatrix} \boldsymbol{\alpha}_\Lambda + \mathbf{z} \right)$$

$$= \frac{1}{m} \sum_{i=1}^m \left(\boldsymbol{\alpha}_\Lambda + \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle} \right) = \boldsymbol{\alpha}_\Lambda + \frac{1}{m} \sum_{i=1}^m \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}.$$

Using this, and since $\sum_{i=1}^m z_i \sim \mathcal{N}(0, m\sigma^2)$, we find

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \mathbb{E}\|\mathbf{H}^*(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\|_2^2$$

11

$$= \mathbb{E}\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2^2 = \mathbb{E}|\widehat{\boldsymbol{\alpha}}_\Lambda - \boldsymbol{\alpha}_\Lambda|^2$$

$$= \mathbb{E}\left|\boldsymbol{\alpha}_\Lambda - \left(\boldsymbol{\alpha}_\Lambda + \frac{1}{m}\sum_{i=1}^m \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}\right)\right|^2$$

$$= \mathbb{E}\left|\frac{1}{m}\sum_{i=1}^m \frac{z_i}{\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle}\right|^2$$

$$= \left(\frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|}\right)^2 \mathbb{E}\left|\sum_{i=1}^m z_i\right|^2$$

$$= \left(\frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|}\right)^2 m\sigma^2 = \frac{1/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2}\sigma^2. \quad (31)$$

Applying the bounds from Lemma 3 to (31), we arrive at

$$\frac{\sigma^2}{m} \leq \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{2m},$$

as desired. □

Corollary 1. *Suppose* $\mathbf{x} = \mathbf{H}^\star \boldsymbol{\alpha}$ *is 1-sparse. Suppose after* $\frac{m}{2}$ *nonadaptive DFT measurements, the support* Λ *is correctly identified. For the remaining* $\frac{m}{2}$ *DFT measurements, we measure repeatedly with a particular measurement from the* $n \times n$ *DFT* $\mathbf{F}$; *denote this measurement by* $\mathbf{f}_j$, *where* $j \in \{0, 1, \ldots, n-1\}$ *is some row index. Then our observations are of the form (20) for* $i = \frac{m}{2} + 1, \ldots, m$, *where the noise* z_i *are i.i.d.* $N(0, \sigma^2)$. *Then the MSE is given by*

$$\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 = \frac{2/m}{|\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|^2}\sigma^2, \quad (32)$$

and is bounded by

$$\frac{2\sigma^2}{m} \leq \mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2 \leq \frac{n\sigma^2}{m}, \quad (33)$$

where the expectation is taken with respect to $\mathbf{z}$.

The upper bound on $\mathbb{E}\|\widehat{\mathbf{x}} - \mathbf{x}\|_2^2$ in Corollary 1 is precisely the lower bound from Theorem 1 when $s = 1$. This means that there is indeed some room for improvement with adaptive sensing when the sparsity basis is the Haar wavelet transform rather than the canonical basis. Corollary 1 shows that the performance of adaptive sensing, in terms of the MSE, depends on the support location of 1-sparse signals. The best adaptive recovery is possible when the support is located on the lowest wavelet frequency ($\Lambda = \{0\}$, or the first Haar wavelet coefficient) while the worst recovery occurs when the support is located on any of the higher wavelet frequencies in block $a = 1$ (the latter half of the Haar wavelet coefficients). This of course matches the intuition based on the correlations in these two bases. This suggests that structured signals such as those that are tree-sparse will benefit more from adaptivity than signals that have a uniformly distributed support.

In light of the discrepancy between (27) and (28), one wishes to know in some sense, what fraction of signals allow for recovery more like one versus the other. Figure 7 shows how $\max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ varies by maximizing $\max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ over Λ while successively removing blocks from $\mathbf{H}^*$. Using our notation for blocks, the blocks of $\mathbf{H}^*$ are removed in the following (top-down) order: $\log_2(n), \log_2(n) - 1, \ldots, 1$. Then, we plot the value
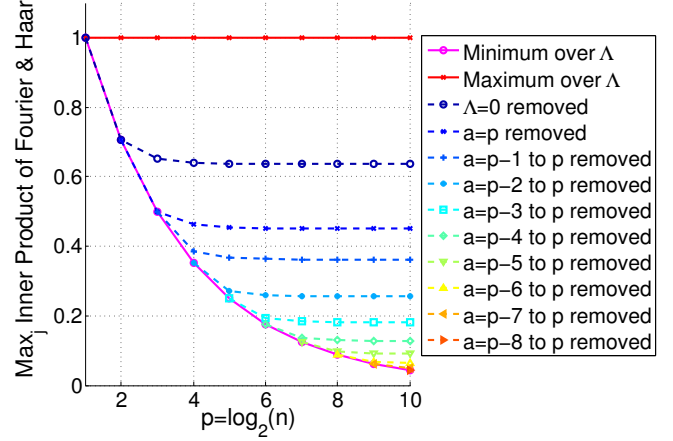

Fig. 7. The value of $\max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ is displayed against the (log of the) signal dimension $n = 2^p$. The solid magenta curve shows the optimization when minimizing over all possible supports $\Lambda \in \{0, \ldots, n-1\}$, given by (27). The solid red curve shows the opposite optimization when maximizing over all possible supports $\Lambda \in \{0, \ldots, n-1\}$, given by (28). The remaining dashed curves show the optimization when maximizing over all supports Λ except those in the blocks indicated.

of $\max_{j \in \{0, \ldots, n-1\}} |\langle \mathbf{f}_j, \mathbf{H}_\Lambda^* \rangle|$ for the remaining submatrix of $\mathbf{H}^*$. Hence, we see that the MSE is higher for signals supported on higher wavelet frequencies, and the upper bound of (33) is achieved by exactly half of the possible signal support sets, whereas the lower bound of (33) is achieved by exactly one of the possible signal supports sets (i.e., $\Lambda = \{0\}$). Fortunately, the support of natural images tends to be concentrated on lower-frequency wavelet coefficients [37].

V. DISCUSSION

Adaptive sensing has tremendous potential to improve the accuracy of sparse recovery in a variety of settings. However, in many practical applications one does not have the freedom to choose arbitrary measurement vectors, but instead must choose from a specified pool of measurements. One example of particular interest is the setting where measurements must be taken from the Fourier ensemble, as is the case in many medical imaging applications. In this paper we established fundamental limitations on the improvements offered by adaptivity in this setting for certain sparsity bases. On the other hand, we argued that for other sparsity bases (such as the Haar wavelet basis) the role of adaptivity in the constrained setting is much less straightforward. We developed a sampling scheme which uses a simple optimization procedure to select measurements adapted to the signal support. This scheme results in significant improvements once an accurate estimate of the support is obtained, which in practice can be achieved by first dedicating a portion of the measurements to support estimation. Though this approach is not necessarily provably optimal, it nonetheless demonstrates the potential of adaptive sensing in the constrained setting. We believe future work in this area can further the understanding of both the limitations of this approach as well as the potential benefits.

REFERENCES

[1] E. Candès and M. Davenport, "How well can we estimate a sparse vector?," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 317–323, 2013.

[2] M. Davenport and E. Arias-Castro, "Compressive binary search," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Cambridge, MA, July 2012.

[3] P. Indyk, E. Price, and D. Woodruff, "On the power of adaptivity in sparse recovery," in *Proc. IEEE Symp. Found. Comp. Science (FOCS)*, Palm Springs, CA, Oct. 2011.

[4] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inform. Theory*, vol. 57, no. 9, pp. 6222–6235, 2011.

[5] M. Malloy and R. Nowak, "Near-optimal adaptive compressed sensing," *IEEE Trans. Inform. Theory*, vol. 60, no. 7, pp. 4001–4012, 2014.

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

[7] E. Candès, "Compressive sampling," in *Proc. Int. Congress of Math.*, Madrid, Spain, Aug. 2006.

[8] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[9] M. Davenport, M. Duarte, Y. Eldar, and G. Kutyniok, "Introduction to compressed sensing," in *Compressed Sensing: Theory and Applications*, Yonina Eldar and Gitta Kutyniok, Eds. Cambridge University Press, Cambridge, UK, 2012.

[10] E. Arias-Castro, E. Candès, and M. Davenport, "On the fundamental limits of adaptive sensing," *IEEE Trans. Inform. Theory*, vol. 59, no. 1, pp. 472–481, 2013.

[11] R. Castro, "Adaptive sensing performance lower bounds for sparse signal detection and support estimation," *Bernoulli*, vol. 20, no. 4, pp. 2217–2246, 2014.

[12] N. Bakhvalov, "On the optimality of linear methods for operator approximation in convex classes of functions," *USSR Computational Mathematics and Mathematical Physics*, vol. 11, no. 4, pp. 244–249, 1971.

[13] H. Woźniakowski, "A survey of information-based complexity," *USSR Computational Mathematics and Mathematical Physics*, vol. 1, no. 1, pp. 11–44, 1985.

[14] E. Novak, "On the power of adaptation," *J. Complexity*, vol. 12, no. 3, pp. 199–237, 1996.

[15] R. Castro, R. Willett, and R. Nowak, "Faster rates in regression via active learning," in *Proc. Adv. in Neural Processing Systems (NIPS)*, 2005, pp. 179–186.

[16] R. Castro, J. Haupt, R. Nowak, and G. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008.

[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[18] H. Nickisch and M. Seeger, "Compressed sensing and Bayesian experimental design," in *Proc. Int. Conf. Machine Learning*, Helsinki, Finland, July 2008.

[19] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Processing Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[20] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, 2009.

[21] H. Hassanieh, P. Indyk, D. Katabi, and E. Price, "Nearly optimal sparse Fourier transform," in *Proc. ACM Symp. Theory of Comput.*, New York, NY, May 2012.

[22] F. Pukelsheim, *Optimal Design of Experiments*, Society for Industrial and Applied Mathematics, 2006.

[23] S. Boyd and L. Vanderberghe, *Convex Optimization*, Cambridge Univ. Press, Cambridge, England, 2004.

[24] I. Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, PA, 1992.

[25] S. Becker, E. Candès, and M. Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Math. Prog. Comput.*, vol. 3, no. 3, pp. 165–218, 2011.

[26] S. Becker, E. Candès, and M. Grant, "Templates for first-order conic solvers user guide, version 1.3 release 2," Tech. Rep., Stanford Univ., Technical report, 2014, Preprint available at http://tfocs.stanford.edu.

[27] M. Duarte, M. Wakin, and R. Baraniuk, "Fast reconstruction of piecewise smooth signals from random projections," in *Proc. Work. Struc. Parc. Rep. Adap. Signaux (SPARS)*, Rennes, France, Nov. 2005.

[28] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[29] F. Krahmer and R. Ward, "Stable and robust sampling strategies for compressive imaging," *IEEE Trans. Image Processing*, vol. 23, no. 2, pp. 612–622, 2014.

[30] F. Krahmer, D. Needell, and R. Ward, "Compressive sensing with redundant dictionaries and structured measurements," *SIAM J. Math. Anal.*, 2015, To appear.

[31] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[32] E. van den Berg and M. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. Sci. Comput.*, vol. 31, no. 2, pp. 890–912, 2008.

[33] E. van den Berg and M. Friedlander, "SPGL1: A solver for large-scale sparse reconstruction," June 2007, http://www.cs.ubc.ca/labs/scl/spgl1.

[34] M. Seeger, H. Nickisch, R. Pohmann, and B. Schölkopf, "Optimization of k-space trajectories for compressed sensing by Bayesian experimental design," *Magnetic Resonance in Medicine*, vol. 63, no. 1, pp. 116–126, 2010.

[35] H. Nickisch, R. Pohmann, B. Schölkopf, and M. Seeger, "Bayesian experimental design of magnetic resonance imaging sequences," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Vancouver, BC, Dec. 2009.

[36] M. Seeger, "Speeding up magnetic resonance image acquisition by Bayesian multi-slice adaptive compressed sensing," in *Proc. Adv. in Neural Processing Systems (NIPS)*, Vancouver, BC, Dec. 2009.

[37] A. Petrosian and F. Meyer, *Wavelets in signal and image analysis: From theory to practice*, Springer Science & Business Media, 2001.

Mark A. Davenport (S'01-M'10-SM'16) is an Assistant Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. Prior to this, he spent 2010-2012 as an NSF Mathematical Sciences Postdoctoral Research Fellow in the Department of Statistics at Stanford University and as a visitor with the Laboratoire Jacques-Louis Lions at the Université Pierre et Marie Curie. He received the B.S.E.E., M.S., and Ph.D. degrees in electrical and computer engineering in 2004, 2007, and 2010, all from Rice University. Prof. Davenport is a recipient of both the National Science Foundation CAREER and Air Force Office of Scientific Research Young Investigator awards. In 2015 he received the Best Paper Award from the IEEE Signal Processing Society. He currently serves as an Associate Editor for *IEEE Transactions on Signal Processing*. His primary area of research concerns the fundamental role that low-dimensional models and optimization play in signal processing, statistical inference, and machine learning.
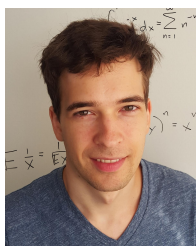
Andrew K. Massimino (S'16) is currently working towards the Ph.D. degree at the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta. He is affiliated with Georgia Techs Center for Signal and Image Processing and received the Algorithms and Randomness Center student fellowship in Spring 2014. He received the B.E. degree in Electrical and Computer Engineering in 2011 from The Cooper Union for the Advancement of Science and Art in New York, New York. His research interests include compressed and adaptive sensing with constraints, recommender systems, and one-bit sensing.
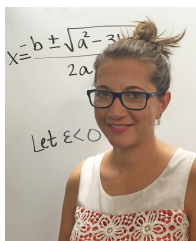
Deanna Needell (M'15) is an Associate Professor in the Mathematical Sciences department at Claremont McKenna College. Prior to this, she spent 2009-2011 as a postdoctoral fellow in the Mathematics and Statistics departments at Stanford University. She recieved her B.S. in Mathematics and Computer Science from the University of Nevada and her M.A. and Ph.D. in Mathematics from the University of California, Davis. Her research interests include Compressed Sensing, Randomized Algorithms, Functional Analysis, Computational Mathematics, Probability, and Statistics. Prof. Needell has been the recipient of awards including an Alfred P. Sloan Fellowship, NSF CAREER award, Simons Foundation Collaboration grant, 2012 IEEE Signal Processing Society Young Author Best Paper Award, and ScienceWatch Fast-Breaking paper award.

Tina Woolf is pursuing the Ph.D. degree in Mathematics at the Institute of Mathematical Sciences, Claremont Graduate University. She received her B.S. in Mathematics from California Polytechnic State University, San Luis Obispo in 2009 and her M.S. in Mathematics from Claremont Graduate University in 2012. Her research interests include compressed sensing, sparse approximation, and stochastic optimization. Since 2009, she has also been a Systems Engineer at Northrop Grumman, primarily involved in modeling, simulation, and payload system performance analysis of infrared systems.