Model-Based Compressive Sensing

Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, Chinmay Hegde

Department of Electrical and Computer Engineering

Rice University

Abstract

Compressive sensing (CS) is an alternative to Shannon/Nyquist sampling for acquisition of sparse or compressible signals that can be well approximated by just $K \ll N$ elements from an N-dimensional basis. Instead of taking periodic samples, we measure inner products with $M < N$ random vectors and then recover the signal via a sparsity-seeking optimization or greedy algorithm. The standard CS theory dictates that robust signal recovery is possible from $M = \mathcal{O}(K \log(N/K))$ measurements. The goal of this paper is to demonstrate that it is possible to substantially decrease M without sacrificing robustness by leveraging more realistic signal models that go beyond simple sparsity and compressibility by including dependencies between values and locations of the signal coefficients. We introduce a model-based CS theory that parallels the conventional theory and provides concrete guidelines on how to create model-based recovery algorithms with provable performance guarantees. A highlight is the introduction of a new class of model-compressible signals along with a new sufficient condition for robust model-compressible signal recovery that we dub the restricted amplification property (RAmP). The RAmP is the natural counterpart to the restricted isometry property (RIP) of conventional CS. To take practical advantage of the new theory, we integrate two relevant signal models — wavelet trees and block sparsity — into two state-of-the-art CS recovery algorithms and prove that they offer robust recovery from just $M = \mathcal{O}(K)$ measurements. Extensive numerical simulations demonstrate the validity and applicability of our new theory and algorithms.

Index Terms

Compressive sensing, sparsity, signal model, union of subspaces, wavelet tree, block sparsity

The authors are listed alphabetically. Email: {richb, volkan, duarte, chinmay}@rice.edu; Web: dsp.rice.edu/cs. This work was supported by the grants NSF CCF-0431150 and CCF-0728867, DARPA/ONR N66001-08-1-2065, ONR N00014-07-1-0936 and N00014-08-1-1112, AFOSR FA9550-07-1-0301, ARO MURI W311NF-07-1-0185, and the Texas Instruments Leadership University Program.

I. Introduction

We are in the midst of a digital revolution that is enabling the development and deployment of new sensors and sensing systems with ever increasing fidelity and resolution. The theoretical foundation is the Shannon/Nyquist sampling theorem, which states that a signal's information is preserved if it is uniformly sampled at a rate at least two times faster than its Fourier bandwidth. Unfortunately, in many important and emerging applications, the resulting Nyquist rate can be so high that we end up with too many samples and must compress in order to store or transmit them. In other applications the cost of signal acquisition is prohibitive, either because of a high cost per sample, or because state-of-the-art samplers cannot achieve the high sampling rates required by Shannon/Nyquist. Examples include radar imaging and exotic imaging modalities outside visible wavelengths.

Transform compression systems reduce the effective dimensionality of an N-dimensional signal x by re-representing it in terms of a sparse set of coefficients α in a basis expansion $x = \Psi\alpha$, with Ψ an $N \times N$ basis matrix. By sparse we mean that only $K \ll N$ of the coefficients α are nonzero and need to be stored or transmitted. By compressible we mean that the coefficients α, when sorted, decay rapidly enough to zero that α can be well-approximated as K-sparse. The sparsity and compressibility properties are pervasive in many signals of interest. For example, smooth signals and images are compressible in the Fourier basis, while piecewise smooth signals and images are compressible in a wavelet basis [1]; the JPEG and JPEG2000 standards are examples of practical transform compression systems based on these bases.

Compressive sensing (CS) provides an alternative to Shannon/Nyquist sampling when the signal under acquisition is known to be sparse or compressible [2–4]. In CS, we measure not periodic signal samples but rather inner products with $M \ll N$ measurement vectors. In matrix notation, the measurements $y = \Phi x = \Phi\Psi\alpha$, where the rows of the $M \times N$ matrix Φ contain the measurement vectors. While the matrix $\Phi\Psi$ is rank deficient, and hence loses information in general, it can be shown to preserve the information in sparse and compressible signals if it satisfies the so-called *restricted isometry property* (RIP) [3]. Intriguingly, a large

class of random matrices have the RIP with high probability. To recover the signal from the compressive measurements y, we search for the sparsest coefficient vector α that agrees with the measurements. To date, research in CS has focused primarily on reducing both the number of measurements M (as a function of N and K) and on increasing the robustness and reducing the computational complexity of the recovery algorithm. Today's state-of-the-art CS systems can robustly recover K-sparse and compressible signals from just $M = \mathcal{O}(K \log(N/K))$ noisy measurements using polynomial-time optimization solvers or greedy algorithms.

While this represents significant progress from Nyquist-rate sampling, our contention in this paper is that it is possible to do even better by more fully leveraging concepts from state-of-the-art signal compression and processing algorithms. In many such algorithms, the key ingredient is a more realistic *signal model* that goes beyond simple sparsity by codifying the inter-dependency *structure* among the signal coefficients α.[1] For instance, JPEG2000 and other modern wavelet image coders exploit not only the fact that most of the wavelet coefficients of a natural image are small but also the fact that the values and locations of the large coefficients have a particular structure. Coding the coefficients according to a model for this structure enables these algorithms to compress images close to the maximum amount possible – significantly better than a naïve coder that just processes each large coefficient independently.

In this paper, we introduce a model-based CS theory that parallels the conventional theory and provides concrete guidelines on how to create model-based recovery algorithms with provable performance guarantees. By reducing the degrees of freedom of a sparse/compressible signal by permitting only certain configurations of the large and zero/small coefficients, signal models provide two immediate benefits to CS. First, they enable us to reduce, in some cases significantly, the number of measurements M required to stably recover a signal. Second, during signal recovery, they enable us to better differentiate true signal information from recovery artifacts, which leads to a more robust recovery.

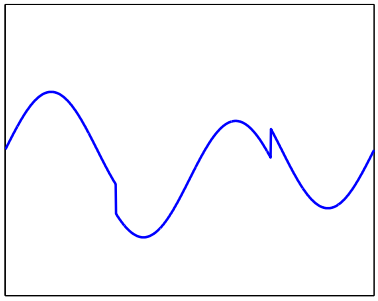
[1]Obviously, sparsity and compressibility correspond to simple signal models where each coefficient is treated independently; for example in a sparse model, the fact that the coefficient α_i is large has no bearing on the size of any α_j, $j \neq i$. We will reserve the use of the term "model" for situations where we are enforcing dependencies between the values and the locations of the coefficients α_i.

To precisely quantify the benefits of model-based CS, we introduce and study several new theoretical concepts that could be of more general interest. We begin with signal models for K-sparse signals and make precise how the structure encoded in a signal model reduces the number of potential sparse signal supports in α. Then using the *model-based restricted isometry property* (RIP) from [5, 6], we prove that such *model-sparse signals* can be robustly recovered from noisy compressive measurements. Moreover, we quantify the required number of measurements M and show that for some models M is independent of N. These results unify and generalize the limited related work to date on signal models for strictly sparse signals [5–9]. We then introduce the notion of a *model-compressible signal*, whose coefficients α are no longer strictly sparse but have a structured power-law decay. To establish that model-compressible signals can be robustly recovered from compressive measurements, we generalize the CS RIP to a new *restricted amplification property* (RAmP). For some compressible signal models, the required number of measurements M is independent of N.

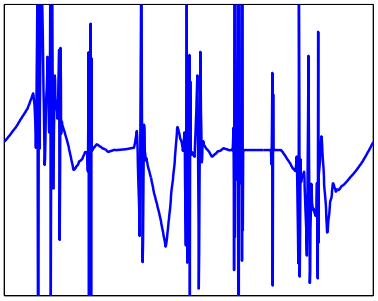
To take practical advantage of this new theory, we demonstrate how to integrate signal models into two state-of-the-art CS recovery algorithms, CoSaMP [10] and iterative hard thresholding (IHT) [11]. The key modification is surprisingly simple: we merely replace the nonlinear approximation step in these greedy algorithms with a model-based approximation. Thanks to our new theory, both new model-based recovery algorithms have provable robustness guarantees for both model-sparse and model-compressible signals.

To validate our theory and algorithms and demonstrate its general applicability and utility, we present two specific instances of model-based CS and conduct a range of simulation experiments. The first model accounts for the fact that the large wavelet coefficients of piecewise smooth signals and images tend to live on a rooted, connected *tree structure* [12]. Using the fact that the number of such trees is much smaller than $\binom{N}{K}$, the number of K-sparse signal supports in N dimensions, we prove that a tree-based CoSaMP algorithm needs only $M = \mathcal{O}(K)$ measurements to robustly recover tree-sparse and tree-compressible signals. Figure 1 indicates the potential performance gains on a tree-compressible, piecewise smooth signal.

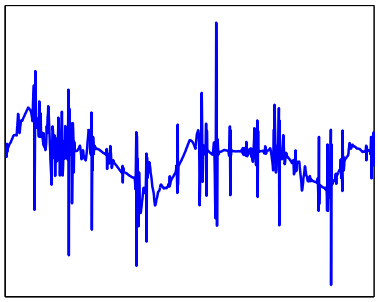
The second model accounts for the fact that the large coefficients of many sparse signals clus-



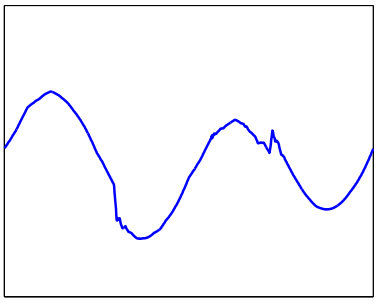
(a) test signal



(b) CoSaMP (RMSE $= 1.123$)



(c) ℓ_1-optimization (RMSE $= 0.751$)



(d) model-based recovery (RMSE $= 0.037$)

Fig. 1. *Example performance of model-based signal recovery. (a) Piecewise-smooth* HeaviSine *test signal of length* $N = 1024$*. This signal is compressible under a connected wavelet tree model. Signal recovered from* $M = 80$ *random Gaussian measurements using (b) the iterative recovery algorithm CoSaMP, (c) standard* ℓ_1 *linear programming, and (d) the wavelet tree-based CoSaMP algorithm from Section V. In all figures, root mean-squared error (RMSE) values are normalized with respect to the* ℓ_2 *norm of the signal.*

ter together [7, 8]. Such a so-called *block sparse* model is equivalent to a *joint sparsity* model for an ensemble of J, length-N signals [9], where the supports of the signals' large coefficients are shared across the ensemble. Using the fact that the number of clustered supports is much smaller than $\binom{JN}{K}$, we prove that a block-based CoSaMP algorithm needs only $M = \mathcal{O}\left(K + \frac{K}{J}\log(\frac{JN}{K})\right)$ measurements to robustly recover block-sparse and block-compressible signals. Moreover, as the number of signals J grows large, the number of measurements approaches $M = \mathcal{O}(K)$.

Our new theory and methods relate to a small body of previous work aimed at integrating signal models with CS. Several groups have developed model-specific signal recovery algorithms [5–8, 13–16]; however, their approach has either been ad hoc or focused on a single model

class. Previous work on unions of subspaces [5, 6, 17] has focused exclusively on strictly sparse signals and has not considered feasible recovery algorithms. To the best of our knowledge, our general framework for model-based recovery, the concept of a model-compressible signal, and the associated RAmP are new to the literature.

This paper is organized as follows. A review of the CS theory in Section II lays out the foundational concepts that we extend to the model-based case in subsequent sections. Section III develops the concept of model-sparse signals and introduces the concept of model-compressible signals. We also quantify how signal models improve the measurement and recovery process by exploiting the model-based RIP for model-sparse signals and by introducing the RAmP for model-compressible signals. Section IV indicates how to tune CoSaMP to incorporate model information and establishes its robustness properties for model-sparse and model-compressible signals. Sections V and VI then specialize our theory to the special cases of wavelet tree and block sparse signal models and report on a series of numerical experiments that validate our theoretical claims. We conclude with a discussion in Section VII. To make the paper more readable, all proofs are relegated to a series of appendices.

II. Background on Compressive Sensing

A. Sparse and compressible signals

Given a basis $\{\psi_i\}_{i=1}^N$, we can represent every signal $x \in \mathbb{R}^N$ in terms of N coefficients $\{\alpha_i\}_{i=1}^N$ as $x = \sum_{i=1}^N \alpha_i \psi_i$; stacking the ψ_i as columns into the $N \times N$ matrix Ψ, we can write succinctly that $x = \Psi\alpha$. In the sequel, we will assume without loss of generality that the signal x is sparse or compressible in the canonical domain so that the sparsity basis Ψ is the identity and $\alpha = x$.

A signal x is K-*sparse* if only $K \ll N$ entries of x are nonzero. We call the set of indices corresponding to the nonzero entries the *support* of x and denote it by $\mathrm{supp}(x)$. The set of all K-sparse signals is the union of the $\binom{N}{K}$, K-dimensional subspaces aligned with the coordinate axes in $\mathbb{R}^N$. We denote this union of subspaces by Σ_K.

Many natural and manmade signals are not strictly sparse, but can be approximated as such;

we call such signals *compressible*. Consider a signal x whose coefficients, when sorted in order of decreasing magnitude, decay according to the power law

$$\left|x_{\mathcal{I}(i)}\right| \leq S\, i^{-1/r}, \quad i = 1, \ldots, N, \tag{1}$$

where $\mathcal{I}$ indexes the sorted coefficients. Thanks to the rapid decay of their coefficients, such signals are well-approximated by K-sparse signals. Let $x_K \in \Sigma_K$ represent the best K-term approximation of x, which is obtained by keeping just the first K terms in $x_{\mathcal{I}(i)}$ from (1). Denote the error of this approximation in the ℓ_p norm as

$$\sigma_K(x)_p := \arg\min_{\bar{x} \in \Sigma_K} \|x - \bar{x}\|_p = \|x - x_K\|_p, \tag{2}$$

where the ℓ_p norm of the vector x is defined as $\|x\|_p = \left(\sum_{i=1}^{N} |x_i|^p\right)^{1/p}$ for $0 < p < \infty$. Then, we have that

$$\sigma_K(x)_p \leq (rs)^{-1/p}\, S K^{-s}, \tag{3}$$

with $s = \frac{1}{r} - \frac{1}{p}$. That is, when measured in the ℓ_p norm, the signal's best approximation error has a power-law decay with exponent s as K increases. We dub such a signal *s-compressible*.

The approximation of compressible signals by sparse signals is the basis of *transform coding* as is used in algorithms like JPEG and JPEG2000 [1]. In this framework, we acquire the full N-sample signal x; compute the complete set of transform coefficients α via $\alpha = \Psi^{-1}x$; locate the K largest coefficients and discard the $(N - K)$ smallest coefficients; and encode the K values and locations of the largest coefficients. While a widely accepted standard, this sample-then-compress framework suffers from three inherent inefficiencies: First, we must start with a potentially large number of samples N even if the ultimate desired K is small. Second, the encoder must compute all of the N transform coefficients α, even though it will discard all but K of them. Third, the encoder faces the overhead of encoding the locations of the large coefficients.

B. Compressive measurements and the restricted isometry property (RIP)

Compressive sensing (CS) integrates the signal acquisition and compression steps into a single process [2–4]. In CS we do not acquire x directly but rather acquire $M < N$ linear

measurements $y = \Phi x$ using an $M \times N$ measurement matrix Φ. We then recover x by exploiting its sparsity or compressibility. Our goal is to push M as close as possible to K in order to perform as much signal "compression" during acquisition as possible.

In order to recover a good estimate of x (the K largest x_i's, for example) from the M compressive measurements, the measurement matrix Φ should satisfy the *restricted isometry property* (RIP) [3].

Definition 1: An $M \times N$ matrix Φ has the *K-restricted isometry property* (K-RIP) with constant δ_K if, for all $x \in \Sigma_K$,

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2. \tag{4}$$

In words, the K-RIP ensures that all submatrices of Φ of size $M \times K$ are close to an isometry, and therefore distance (and information) preserving. Practical recovery algorithms typically require that Φ have a slightly stronger $2K$-RIP, $3K$-RIP, or higher-order RIP in order to preserve distances between K-sparse vectors (which are $2K$-sparse in general), three-way sums of K-sparse vectors (which are $3K$-sparse in general), and other higher-order structures.

While the design of a measurement matrix Φ satisfying the K-RIP is an NP-Complete problem in general [3], random matrices whose entries are i.i.d. Gaussian, Bernoulli (± 1), or more generally subgaussian[2] work with high probability provided $M = \mathcal{O}\left(K \log(N/K)\right)$. These random matrices also have a so-called *universality* property in that, for any choice of orthonormal basis matrix Ψ, $\Phi\Psi$ has the K-RIP with high probability. This is useful when the signal is sparse not in the canonical domain but in basis Ψ. A random Φ corresponds to an intriguing data acquisition protocol in which each measurement y_j is a randomly weighted linear combination of the entries of x.

[2]A random variable X is called subgaussian if there exists $c > 0$ such that $\mathbb{E}\left(e^{Xt}\right) \leq e^{c^2 t^2/2}$ for all $t \in \mathbb{R}$. Examples include the Gaussian and Bernoulli random variables, as well as any bounded random variable. [18]

C. Recovery algorithms

Since there are infinitely many signal coefficient vectors x' that produce the same set of compressive measurements $y = \Phi x$, to recover the "right" signal we exploit our a priori knowledge of its sparsity or compressibility. For example, we could seek the sparsest x that agrees with the measurements y:

$$x = \arg\min_{y=\Phi x'} \|x'\|_0, \tag{5}$$

where the ℓ_0 "norm" of a vector counts its number of nonzero entries. While this optimization can recover a K-sparse signal from just $M = 2K$ compressive measurements, it is unfortunately a combinatorial, NP-Complete problem; furthermore, the recovery is not stable in the presence of noise.

Practical, stable recovery algorithms rely on the RIP (and therefore require at least $M = \mathcal{O}(K \log(N/K))$ measurements); they can be grouped into two camps. The first approach convexifies the ℓ_0 optimization (5) to the ℓ_1 optimization

$$x = \arg\min_{y=\Phi x'} \|x'\|_1. \tag{6}$$

This corresponds to a linear program that can be solved in polynomial time [2, 3]. Adaptations to deal with additive noise in y or x include basis pursuit with denoising (BPDN) [19], complexity-based regularization [20], and the Dantzig Selector [21].

The second approach finds the sparsest x agreeing with the measurements y through an iterative, greedy search. Algorithms such as matching pursuit, orthogonal matching pursuit [22], StOMP [23], iterative hard thresholding (IHT) [11], CoSaMP [10], and Subspace Pursuit (SP) [24] all revolve around a best L-term approximation for the estimated signal, with L varying for each algorithm.

D. Performance bounds on signal recovery

Given $M = \mathcal{O}(K \log(N/K))$ compressive measurements, a number of different CS signal recovery algorithms, including all of the ℓ_1 techniques mentioned above and the CoSaMP, SP,

and IHT iterative techniques, offer provably stable signal recovery with performance close to optimal K-term approximation (recall (3)). For a random Φ, all results hold with high probability.

For a noise-free, K-sparse signal, these algorithms offer perfect recovery, meaning that the signal $\widehat{x}$ recovered from the compressive measurements $y = \Phi x$ is exactly $\widehat{x} = x$.

For a K-sparse signal x whose measurements are corrupted by noise n of bounded norm — that is, we measure $y = \Phi x + n$ — the mean-squared error of the recovered signal $\widehat{x}$ is

$$\|x - \widehat{x}\|_2 \leq C\|n\|_2, \tag{7}$$

with C a small constant [2, 3, 10, 11].

For an s-compressible signal x whose measurements are corrupted by noise n of bounded norm, the mean-squared error of the recovered signal $\widehat{x}$ is

$$\|x - \widehat{x}\|_2 \leq C_1\|x - x_K\|_2 + C_2\frac{1}{\sqrt{K}}\|x - x_K\|_1 + C_3\|n\|_2. \tag{8}$$

Using (3) we can simplify this expression to

$$\|x - \widehat{x}\|_2 \leq \frac{C_1 S K^{-s}}{\sqrt{2s}} + \frac{C_2 S K^{-s}}{s - 1/2} + C_3\|n\|_2. \tag{9}$$

III. Beyond Sparse and Compressible Signals

While many natural and manmade signals and images can be described to first-order as sparse or compressible, the support of their large coefficients often has an underlying inter-dependency structure. This phenomenon has received only limited attention by the CS community to date [5–8, 14–16]. In this section, we introduce a model-based theory of CS that captures such structure. A model reduces the degrees of freedom of a sparse/compressible signal by permitting only certain configurations of supports for the large coefficient. As we will show, this allows us to reduce, in some cases significantly, the number of compressive measurements M required to stably recover a signal.

A. Model-sparse signals

Recall from Section II-A that a K-sparse signal vector x lives in $\Sigma_K \subset \mathbb{R}^N$, which is a union of $\binom{N}{K}$ subspaces of dimension K. Other than its K-sparsity, there are no further constraints on the support or values of its coefficients. A *signal model* endows the K-sparse signal x with additional structure that allows certain K-dimensional subspaces in Σ_K and disallows others [5, 6].

To state a formal definition of a signal model, let $x|_\Omega$ represent the entries of x corresponding to the set of indices $\Omega \subseteq \{1, \ldots, N\}$, and let Ω^C denote the complement of the set Ω.

Definition 2: A *signal model* $\mathcal{M}_K$ is defined as the union of m_K canonical K-dimensional subspaces

$$\mathcal{M}_K = \bigcup_{m=1}^{m_K} \mathcal{X}_m, \text{ such that } \mathcal{X}_m := \{x : x|_{\Omega_m} \in \mathbb{R}^K, x|_{\Omega_m^C} = 0\},$$

where each subspace $\mathcal{X}_m$ contains all signals x with $\mathrm{supp}(x) \in \Omega_m$. Thus, the model $\mathcal{M}_K$ is defined by the set of possible supports $\{\Omega_1, \ldots, \Omega_{m_K}\}$.

Signals from $\mathcal{M}_K$ are called *K-model sparse*. Clearly, $\mathcal{M}_K \subseteq \Sigma_K$ and contains $m_K \leq \binom{N}{K}$ subspaces.

In Sections V and VI below we consider two concrete models for sparse signals. The first model accounts for the fact that the large wavelet coefficients of piecewise smooth signals and images tend to live on a rooted, connected *tree structure* [12]. The second model accounts for the fact that the large coefficients of sparse signals often *cluster* together [7–9].

B. Model-based RIP

If we know that the signal x being acquired is K-model sparse, then we can relax the RIP constraint on the CS measurement matrix Φ and still achieve stable recovery from the compressive measurements $y = \Phi x$ [5, 6].

Definition 3: [5, 6] An $M \times N$ matrix Φ has the *$\mathcal{M}_K$-restricted isometry property* ($\mathcal{M}_K$-RIP) with constant $\delta_{\mathcal{M}_K}$ if, for all $x \in \mathcal{M}_K$, we have

$$(1 - \delta_{\mathcal{M}_K})\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_{\mathcal{M}_K})\|x\|_2^2. \tag{10}$$

To obtain a performance guarantee for model-based recovery of K-model sparse signals in additive measurement noise, we must define an enlarged union of subspaces that includes sums of elements in the model.

Definition 4: The *B-Minkowski sum* for the set $\mathcal{M}_K$, with $B > 1$ an integer, is defined as

$$\mathcal{M}_K^B = \left\{ x = \sum_{r=1}^{B} x^{(r)}, \text{ with } x^{(r)} \in \mathcal{M}_K \right\}.$$

Define $\mathbb{M}_B(x, K)$ as the algorithm that obtains the best approximation of x in the enlarged union of subspaces $\mathcal{M}_K^B$:

$$\mathbb{M}_B(x, K) = \arg \min_{\bar{x} \in \mathcal{M}_K^B} \|x - \bar{x}\|_2.$$

We write $\mathbb{M}(x, K) := \mathbb{M}_1(x, K)$ when $B = 1$. Note that for many models, we will have $\mathcal{M}_K^B \subset \mathcal{M}_{BK}$, and so the algorithm $\mathbb{M}(x, BK)$ will provide a strictly better approximation than $\mathbb{M}_B(x, K)$.

Our performance guarantee for model-sparse signal recovery will require that the measurement matrix Φ be a near-isometry for all subspaces in $\mathcal{M}_K^B$ for some $B > 1$. This requirement is a direct generalization of the $2K$-RIP, $3K$-RIP, and higher-order RIPs from the conventional CS theory.

Blumensath and Davies [5] have quantified the number of measurements M necessary for a random CS matrix to have the $\mathcal{M}_K$-RIP with a given probability.

Theorem 1: [5] Let $\mathcal{M}_K$ be the union of m_K subspaces of K-dimensions in $\mathbb{R}^N$. Then, for any $t > 0$ and any

$$M \geq \frac{2}{c\delta_{\mathcal{M}_K}^2} \left(\ln(2m_K) + K \ln \frac{12}{\delta_{\mathcal{M}_K}} + t \right),$$

an $M \times N$ i.i.d. subgaussian random matrix has the $\mathcal{M}_K$-RIP with constant $\delta_{\mathcal{M}_K}$ with probability at least $1 - e^{-t}$.

This bound can be used to recover the conventional CS result by substituting $m_K = \binom{N}{K} \approx (Ne/K)^K$. The $\mathcal{M}_K$-RIP property is sufficient for robust recovery of model-sparse signals, as we show below in Section IV-B.

C. Model-compressible signals

Just as compressible signals are "nearly K-sparse" and thus live close to the union of subspaces Σ_K in $\mathbb{R}^N$, model-compressible signals are "nearly K-model sparse" and live close to the restricted union of subspaces $\mathcal{M}_K$. In this section, we make this new concept rigorous. Recall from (3) that we defined compressible signals in terms of the decay of their K-term approximation error.

The ℓ_2 error incurred by approximating $x \in \mathbb{R}^N$ by the best model-based approximation in $\mathcal{M}_K$ is given by

$$\sigma_{\mathcal{M}_K}(x) := \inf_{\bar{x} \in \mathcal{M}_K} \|x - \bar{x}\|_2 = \|x - \mathbb{M}(x, K)\|_2.$$

The decay of this approximation error defines the model-compressibility of a signal.

Definition 5: The set of *s-model-compressible signals* is defined as

$$\mathfrak{M}_s = \left\{ x \in \mathbb{R}^N : \sigma_{\mathcal{M}_K}(x) \leq S K^{-1/s}, 1 \leq K \leq N, S < \infty \right\}.$$

Define $|x|_{\mathfrak{M}_s}$ as the smallest value of S for which this condition holds for x and s.

We say that $x \in \mathfrak{M}_s$ is an *s-model-compressible signal* under the signal model $\mathcal{M}_K$. These approximation classes have been characterized for certain signal models; see Section V for an example.

D. Nested model approximations and residual subspaces

In conventional CS, the same requirement (RIP) is a sufficient condition for the stable recovery of both sparse and compressible signals. In model-based recovery, however, the class of compressible signals is much larger than that of sparse signals, since the set of subspaces containing model-sparse signals does not span all K-dimensional subspaces. Therefore, we need to introduce some additional tools to develop a sufficient condition for the stable recovery of model-compressible signals.

We will pay particular attention to models $\mathcal{M}_K$ that generate *nested approximations*, since they are more amenable to analysis.

Definition 6: A model $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots\}$ has the *nested approximation property* (NAP) if $\text{supp}(\mathbb{M}(x, K)) \subset \text{supp}(\mathbb{M}(x, K'))$ for all $K < K'$ and for all $x \in \mathbb{R}^N$.

In words, a model generates nested approximations if the support of the best K'-term model-based approximation contains the support of the best K-term model-based approximation for all $K < K'$. An important example of a NAP model is the standard compressible signal model of (3).

When a model obeys the NAP, the support of the difference between the best jK-term model-based approximation and the best $(j+1)K$-term model-based approximation of a signal can be shown to lie in a small union of subspaces, thanks to the structure enforced by the model. This structure is captured by the set of subspaces that are included in each subsequent approximation, as defined below.

Definition 7: The jth set of *residual subspaces* of size K is defined as

$$\mathcal{R}_{j,K}(\mathcal{M}) = \left\{u \in \mathbb{R}^N \text{ such that } u = \mathbb{M}(x, jK) - \mathbb{M}(x, (j-1)K) \text{ for some } x \in \mathbb{R}^N\right\},$$

for $j = 1, \ldots, \lceil N/K \rceil$.

Under the NAP, each signal x in a model can be partitioned into its best K-term approximation x_{T_1}, the additional components present in the best $2K$-term approximation x_{T_2}, and so on, with $x = \sum_{j=1}^{\lceil N/K \rceil} x_{T_j}$ and $x_{T_j} \in \mathcal{R}_{j,K}(\mathcal{M})$ for each j. Each signal partition x_{T_j} is a K-sparse signal, and thus $\mathcal{R}_{j,K}(\mathcal{M})$ is a union of subspaces of dimension K. We will denote by R_j the number of subspaces that compose $\mathcal{R}_{j,K}(\mathcal{M})$ and omit the dependence on $\mathcal{M}$ in the sequel for brevity.

Intuitively, the norms of the partitions $\|x_{T_j}\|_2$ decay as j increase for signals that are compressible under the model. As the next subsection shows, this observation is instrumental in relaxing the isometry restrictions on the measurement matrix Φ and bounding the recovery error for s-model-compressible signals when the model obeys the NAP.

E. The restricted amplification property (RAmP)

For exactly K-model-sparse signals, we discussed in Section III-B that the number of compressive measurements M required for a random matrix to have the $\mathcal{M}_K$-RIP is determined by the number of canonical subspaces m_K via (11). Unfortunately, such model-sparse concepts and results do not immediately extend to model-compressible signals. Thus, we develop a generalization of the $\mathcal{M}_K$-RIP that we will use to quantify the stability of recovery for model-compressible signals.

One way to analyze the robustness of compressible signal recovery in conventional CS is to consider the tail of the signal outside its K-term approximation as contributing additional "noise" to the measurements of size $\|\Phi(x - x_K)\|_2$ [10, 11, 25]. Consequently, the conventional K-sparse recovery performance result can be applied with the augmented noise $n + \Phi(x - x_K)$.

This technique can also be used to quantify the robustness of model-compressible signal recovery. The key quantity we must control is the amplification of the model-based approximation residual through Φ. The following property is a new generalization of the RIP and model-based RIP.

Definition 8: A matrix Φ has the (ϵ_K, r)*-restricted amplification property* (RAmP) for the residual subspaces $\mathcal{R}_{j,K}$ of model $\mathcal{M}$ if

$$\|\Phi u\|_2^2 \leq (1 + \epsilon_K) j^{2r} \|u\|_2^2 \tag{11}$$

for any $u \in \mathcal{R}_{j,K}$ for each $1 \leq j \leq \lceil N/K \rceil$.

The regularity parameter $r > 0$ caps the growth rate of the amplification of $u \in \mathcal{R}_{j,K}$ as a function of j. Its value can be chosen so that the growth in amplification with j balances the decay of the norm in each residual subspace $\mathcal{R}_{j,K}$ with j.

We can quantify the number of compressive measurements M required for a random measurement matrix Φ to have the RAmP with high probability; we prove the following in Appendix I.

Theorem 2: Let Φ be an $M \times N$ matrix with i.i.d. subgaussian entries and let the set of residual subspaces $\mathcal{R}_{j,K}$ of model $\mathcal{M}$ contain R_j subspaces of dimension K for each $1 \leq j \leq$

$\lceil N/K \rceil$. If

$$M \geq \max_{1 \leq j \leq \lceil N/K \rceil} \frac{1}{\left(j^r \sqrt{1+\epsilon_K} - 1\right)^2} \left(2K + 4 \ln \frac{R_j N}{K} + 2t\right), \tag{12}$$

then the matrix Φ has the (ϵ_K, r)-RAmP with probability $1 - e^{-t}$.

The order of the bound of Theorem 2 is lower than $\mathcal{O}\left(K \log(N/K)\right)$ as long as the number of subspaces R_j grows slower than N^K.

Armed with the RaMP, we can state the following result, which will provide robustness for the recovery of model-compressible signals; see Appendix II for the proof.

Theorem 3: Let $x \in \mathfrak{M}_s$ be an s-model compressible signal under a model $\mathcal{M}$ that obeys the NAP. If Φ has the (ϵ_K, r)-RAmP and $r = s - 1$, then we have

$$\|\Phi(x - \mathbb{M}(x, K))\|_2 \leq \sqrt{1+\epsilon_K} K^{-s} \ln \left\lceil \frac{N}{K} \right\rceil |x|_{\mathfrak{M}_s}.$$

IV. Model-Based Signal Recovery Algorithms

To take practical advantage of our new theory for model-based CS, we demonstrate how to integrate signal models into two state-of-the-art CS recovery algorithms, CoSaMP [10] (in this section) and iterative hard thresholding (IHT) [11] (in Appendix III). The key modification is simple: we merely replace the best K-term approximation step in these greedy algorithms with a best K-term model-based approximation. Since at each iteration we need only search over the m_K subspaces of $\mathcal{M}_K$ rather than $\binom{N}{K}$ subspaces of Σ_K, fewer measurements will be required for the same degree of robust signal recovery. Or, alternatively, using the same number of measurements, more accurate recovery can be achieved. After presenting the modified CoSaMP algorithm, we prove robustness guarantees for both model-sparse and model-compressible signals.

A. Model-based CoSaMP

We choose to modify the CoSaMP algorithm [10] for two reasons. First, it has robust recovery guarantees that are on par with the best convex optimization-based approaches. Second,

Algorithm 1 Model-based CoSaMP

Inputs: CS matrix Φ, measurements y, model $\mathcal{M}_K$

Output: K-sparse approximation $\widehat{x}$ to true signal x

$\widehat{x}_0 = 0$, $r = y$, $i = 0$ {initialize}

while halting criterion false **do**

  $i \leftarrow i + 1$

  $e \leftarrow \Phi^T r$ {form signal residual estimate}

  $\Omega \leftarrow \mathrm{supp}(\mathbb{M}_2(e, K))$ {prune signal residual estimate according to signal model}

  $T \leftarrow \Omega \cup \mathrm{supp}(\widehat{x}_{i-1})$ {merge supports}

  $b|_T \leftarrow \Phi_T^{\dagger} y$, $b|_{T^C} \leftarrow 0$ {form signal estimate}

  $\widehat{x}_i \leftarrow \mathbb{M}(b, K)$ {prune signal estimate according to signal model}

  $r \leftarrow y - \Phi \widehat{x}_i$ {update measurement residual}

end while

return $\widehat{x} \leftarrow \widehat{x}_i$

it has a simple iterative, greedy structure based on a best BK-term approximation (with B a small integer) that is easily modified to incorporate a best BK-term model-based approximation $\mathbb{M}_B(K, x)$. Pseudocode for the modified algorithm is given in Algorithm 1.

We now study the performance of model-based CoSaMP signal recovery on model-sparse signals and model-compressible signals.

B. Performance of model-sparse signal recovery

A robustness guarantee for noisy measurements of model-sparse signals can be obtained using the model-based RIP (10). The following theorem is proven in Appendix IV.

Theorem 4: Let $x \in \mathcal{M}_K$ and let $y = \Phi x + n$ be a set of noisy CS measurements. If Φ has an $\mathcal{M}_K^4$-RIP constant of $\delta_{\mathcal{M}_K^4} \leq 0.1$, then the signal estimate $\widehat{x}_i$ obtained from iteration i of the model-based CoSaMP algorithm satisfies

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 15\|n\|_2. \tag{13}$$

C. Performance of model-compressible signal recovery

Using the new tools introduced in Section III, we can provide a robustness guarantee for noisy measurements of model-compressible signals, using the RAmP as a condition on the measurement matrix Φ.

Theorem 5: Let $x \in \mathfrak{M}_s$ be an s-model-compressible signal from a model $\mathcal{M}$ that obeys the NAP, and let $y = \Phi x + n$ be a set of noisy CS measurements. If Φ has the $\mathcal{M}_K^4$-RIP with $\delta_{\mathcal{M}_K^4} \leq 0.1$ and the (ϵ_K, r)-RAmP with $\epsilon_K \leq 0.1$ and $r = s - 1$, then the signal estimate $\widehat{x}_i$ obtained from iteration i of the model-based CoSaMP algorithm satisfies

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 35\left(\|n\|_2 + |x|_{\mathfrak{M}_s} K^{-s}(1 + \ln\lceil N/K \rceil)\right). \tag{14}$$

To prove the theorem, we first bound the recovery error for an s-model-compressible signal $x \in \mathfrak{M}_s$ when the matrix Φ has the (ϵ_K, r)-RAmP with $r \leq s - 1$. Then, using Theorems 3 and 4, we can easily prove the result by following the analogous proof in [10].

D. Robustness to model mismatch

We now analyze the robustness of model-based CS recovery to *model mismatch*, which occurs when the signal being recovered from compressive measurements does not conform exactly to the model used in the recovery algorithm.

We begin with optimistic results for signals that are "close" to matching the recovery model. First consider a signal x that is not K-model sparse as the recovery algorithm assumes but rather $(K + \kappa)$-model sparse for some small integer κ. This signal can be decomposed into x_K, the signal's K-term model-based approximation, and $x - x_K$, the error of this approximation. For $\kappa \leq K$, we have that $x - x_K \in \mathcal{R}_{2,K}$. If the matrix Φ has the (ϵ_K, r)-RAmP, then it follows than

$$\|\Phi(x - x_K)\|_2 \leq 2^r \sqrt{1 + \epsilon_K}\|x - x_K\|_2. \tag{15}$$

Using equations (13) and (15), we obtain the following guarantee for the ith iteration of model-based CoSaMP:

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 16 \cdot 2^r \sqrt{1+\epsilon_K}\|x - x_K\|_2 + 15\|n\|_2.$$

By noting that $\|x - x_K\|_2$ is small, we obtain a guarantee that is close to (13).

Second, consider a signal x that is not s-model compressible as the recovery algorithm assumes but rather $(s-\epsilon)$-model compressible. The following bound can be obtained under the conditions of Theorem 5 by modifying the argument in Appendix II:

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 35\left(\|n\|_2 + |x|_{\mathfrak{M}_s} K^{-s}\left(1 + \frac{\lceil N/K\rceil^{\epsilon} - 1}{\epsilon}\right)\right).$$

As ϵ becomes smaller, the factor $\frac{\lceil N/K\rceil^{\epsilon}-1}{\epsilon}$ approaches $\log\lceil N/K\rceil$, matching (14). In summary, as long as the deviations from the model-sparse and model-compressible models are small, our model-based recovery guarantees still apply within a small bounded constant factor.

We end with a more pessimistic, worst-case result for signals that are arbitrarily far away from model-sparse or model-compressible. Consider such an arbitrary $x \in \mathbb{R}^N$ and compute its nested model-based approximations $x_{jK} = \mathbb{M}(x, jK)$, $j = 1, \ldots, \lceil N/K\rceil$. If x is not model-compressible, then the model-based approximation error $\sigma_{jK}(x)$ is not guaranteed to decay as j decreases. Additionally, the number of residual subspaces $\mathcal{R}_{j,K}$ could be as large as $\binom{N}{K}$; that is, the jth difference between subsequent model-based approximations $x_{T_j} = x_{jK} - x_{(j-1)K}$ might lie in any arbitrary K-dimensional subspace. This worst case is equivalent to setting $r = 0$ and $R_j = \binom{N}{K}$ in Theorem 2. It is easy to see that this condition on the number of measurements M is nothing but the standard RIP for CS. Hence, if inflate the number of measurements to $M = \mathcal{O}\left(K\log(N/K)\right)$ (the usual number for conventional CS), the performance of model-based CoSaMP recovery on an arbitrary signal x follows the K-term *model-based* approximation of x within a bounded constant factor.

E. Computational complexity of model-based recovery

The computational complexity of a model-based signal recovery algorithm differs from that of a standard algorithm by two factors. The first factor is the reduction in the number

of measurements M necessary for recovery: since most current recovery algorithms have a computational complexity that is linear in the number of measurements, any reduction in M reduces the total complexity. The second factor is the cost of the model-based approximation. The K-term approximation used in most current recovery algorithms can be implemented with a simple sorting operation ($\mathcal{O}(N \log N)$ complexity, in general). Ideally, the signal model should support a similarly efficient approximation algorithm.

To validate our theory and algorithms and demonstrate their general applicability and utility, we now present two specific instances of model-based CS and conduct a range of simulation experiments.

V. Example: Wavelet Tree Model

Wavelet decompositions have found wide application in the analysis, processing, and compression of smooth and piecewise smooth signals because these signals are K-sparse and compressible, respectively [1]. Moreover, the wavelet coefficients can be naturally organized into a tree structure, and for many kinds of natural and manmade signals the largest coefficients cluster along the branches of this tree. This motivates a connected tree model for the wavelet coefficients [26–28].

While CS recovery for wavelet-sparse signals has been considered previously [14–16], the resulting algorithms integrated the tree constraint in an ad-hoc fashion. Furthermore, the algorithms provide no recovery guarantees or bounds on the necessary number of compressive measurements.

A. *Tree-sparse signals*

We first describe tree sparsity in the context of sparse wavelet decompositions. We focus on one-dimensional signals and binary wavelet trees, but all of our results extend directly to d-dimensional signals and 2^d-ary wavelet trees.

Consider a signal x of length $N = 2^I$, for an integer value of I. The wavelet representation

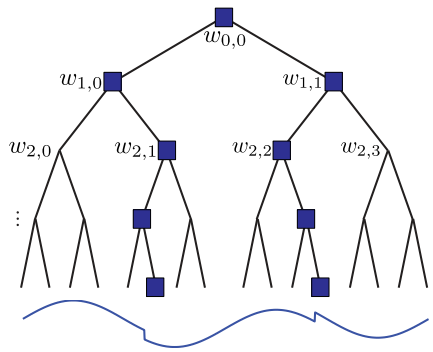


Fig. 2. *Binary wavelet tree for a one-dimensional signal. The squares denote the large wavelet coefficients that arise from the discontinuities in the piecewise smooth signal drawn below; the support of the large coefficients forms a rooted, connected tree.*

of x is given by

$$x = v_0 \nu + \sum_{i=0}^{I-1} \sum_{j=0}^{2^i-1} w_{i,j} \psi_{i,j},$$

where ν is the scaling function and $\psi_{i,j}$ is the wavelet function at scale i and offset j. The wavelet transform consists of the scaling coefficient v_0 and wavelet coefficients $w_{i,j}$ at scale i, $0 \leq i \leq I-1$, and position j, $0 \leq j \leq 2^i - 1$. In terms of our earlier matrix notation, x has the representation $x = \Psi\alpha$, where Ψ is a matrix containing the scaling and wavelet functions as columns, and $\alpha = [v_0 \; w_{0,0} \; w_{1,0} \; w_{1,1} \; w_{2,0} \ldots]^T$ is the vector of scaling and wavelet coefficients. We are, of course, interested in sparse and compressible α.

The nested supports of the wavelets at different scales create a parent/child relationship between wavelet coefficients at different scales. We say that $w_{i-1,\lfloor j/2 \rfloor}$ is the *parent* of $w_{i,j}$ and that $w_{i+1,2j}$ and $w_{i+1,2j+1}$ are the *children* of $w_{i,j}$. These relationships can be expressed graphically by the wavelet coefficient tree in Figure 2.

Wavelet functions act as local discontinuity detectors, and using the nested support property of wavelets at different scales, it is straightforward to see that a signal discontinuity will give rise to a chain of large wavelet coefficients along a branch of the wavelet tree from a leaf to the root. Moreover, smooth signal regions will give rise to regions of small wavelet coefficients. This "connected tree" property has been well-exploited in a number of wavelet-based processing

[12, 29, 30] and compression [31, 32] algorithms. In this section, we will specialize the theory developed in Sections III and IV to a connected tree model $\mathcal{T}$.

A set of wavelet coefficients Ω forms a *connected subtree* if, whenever a coefficient $w_{i,j} \in \Omega$, then its parent $w_{i-1,\lfloor j/2 \rfloor} \in \Omega$ as well. Each such set Ω defines a subspace of signals whose support is contained in Ω; that is, all wavelet coefficients outside Ω are zero. In this way, we define the model $\mathcal{T}_K$ as the union of all K-dimensional subspaces corresponding to supports Ω that form connected subtrees.

Definition 9: Define the set of K*-tree sparse signals* as

$$\mathcal{T}_K = \left\{ x = v_0 \nu + \sum_{i=0}^{I-1} \sum_{j=1}^{2^i} w_{i,j} \psi_{i,j} : w|_{\Omega^C} = 0, |\Omega| = K, \Omega \text{ forms a connected subtree} \right\}.$$

To quantify the number of subspaces in $\mathcal{T}_K$, it suffices to count the number of distinct connected subtrees of size K in a binary tree of size N. We prove the following result in Appendix V.

Proposition 1: The number of subspaces in $\mathcal{T}_K$ obeys $T_K \leq \frac{4^{K+4}}{Ke^2}$ for $K \geq \log_2 N$ and $T_K \leq \frac{(2e)^K}{K+1}$ for $K < \log_2 N$.

B. Tree-based approximation

To implement tree-based signal recovery, we seek an efficient algorithm $\mathbb{T}(x, K)$ to solve the optimal approximation

$$x_K^{\mathcal{T}} = \arg \min_{\bar{x} \in \mathcal{T}_K} \|x - \bar{x}\|_2. \tag{16}$$

Fortuitously, an efficient solver exists, called the *condensing sort and select algorithm* (CSSA) [26–28]. Recall that subtree approximation coincides with standard K-term approximation (and hence can be solved by simply sorting the wavelet coefficients) when the wavelet coefficients are monotonically nonincreasing along the tree branches out from the root. The CSSA solves (16) in the case of general wavelet coefficient values by *condensing* the nonmonotonic segments of the tree branches using an iterative sort-and-average routine. The condensed nodes are called "supernodes". Condensing a large coefficient far down the tree accounts for the potentially large cost (in terms of the total budget of tree nodes K) of growing the tree to that point.

The CSSA can also be interpreted as a greedy search among the nodes. For each node in the tree, the algorithm calculates the average wavelet coefficient magnitude for each subtree rooted at that node, and records the largest average among all the subtrees as the energy for that node. The CSSA then searches for the unselected node with the largest energy and adds the subtree corresponding to the node's energy to the estimated support as a supernode [28].

Since the first step of the CSSA involves sorting all of the wavelet coefficients, overall it requires $\mathcal{O}(N \log N)$ computations. However, once the CSSA grows the optimal tree of size K, it is trivial to determine the optimal trees of size $< K$ and computationally efficient to grow the optimal trees of size $> K$ [26].

The constrained optimization (16) can be rewritten as an unconstrained problem by introducing the Lagrange multiplier λ [33]:

$$\min_{\bar{x} \in \bar{\mathcal{T}}} \|x - \bar{x}\|_2^2 + \lambda(\|\bar{\alpha}\|_0 - K),$$

where $\bar{\mathcal{T}} = \cup_{n=1}^{N} \mathcal{T}_n$ and $\bar{\alpha}$ are the wavelet coefficients of $\bar{x}$. Except for the inconsequential λK term, this optimization coincides with Donoho's *complexity penalized sum of squares* [33], which can be solved in only $\mathcal{O}(N)$ computations using coarse-to-fine dynamic programming on the tree. Its primary shortcoming is the nonobvious relationship between the tuning parameter λ and and the resulting size K of the optimal connected subtree.

C. Tree-compressible signals

Specializing Definition 2 from Section III-C to $\mathcal{T}$, we make the following definition.

Definition 10: Define the set of *s-tree compressible signals* as

$$\mathfrak{T}_s = \left\{ x \in \mathbb{R}^N : \|x - \mathbb{T}(x, K)\|_2 \leq SK^{-s}, 1 \leq K \leq N, S < \infty \right\}.$$

Furthermore, define $|x|_{\mathfrak{T}_s}$ as the smallest value of S for which this condition holds for x and s.

Tree approximation classes contain signals whose wavelet coefficients have a loose (and possibly interrupted) decay from coarse to fine scales. These classes have been well-characterized for wavelet-sparse signals [27, 28, 32] and are intrinsically linked with the Besov spaces

$B_q^s(L_p([0,1]))$. Besov spaces contain functions of one or more continuous variables that have (roughly speaking) s derivatives in $L_p([0,1])$; the parameter q provides finer distinctions of smoothness. When a Besov space signal x_a with $s > 1/p - 1/2$ is sampled uniformly and converted to a length-N vector x, its wavelet coefficients belong to the tree approximation space $\mathfrak{T}_s$, with

$$|x_N|_{\mathfrak{T}_s} \asymp \|x_a\|_{L_p([0,1])} + \|x_a\|_{B_q^s(L_p([0,1]))},$$

where "$\asymp$" denotes an equivalent norm. The same result holds if $s = 1/p - 1/2$ and $q \leq p$.

D. Stable tree-based recovery from compressive measurements

For tree-sparse signals, by applying Theorem 1 and Proposition 1, we find that a subgaussian random matrix has the $\mathcal{T}_K$-RIP property with constant $\delta_{\mathcal{T}_K}$ and probability $1 - e^{-t}$ if the number of measurements obeys

$$M \geq \begin{cases} \frac{2}{c\delta_{\mathcal{T}_K}^2}\left(K \ln \frac{48}{\delta_{\mathcal{T}_K}} + \ln \frac{512}{Ke^2} + t\right) & \text{if } K < \log_2 N, \\ \frac{2}{c\delta_{\mathcal{T}_K}^2}\left(K \ln \frac{24e}{\delta_{\mathcal{T}_K}} + \ln \frac{2}{K+1} + t\right) & \text{if } K \geq \log_2 N, \end{cases}$$

Thus, the number of measurements necessary for stable recovery of tree-sparse signals is linear in K, without the dependence on N present in conventional non-model-based CS recovery.

For tree-compressible signals, we must quantify the number of subspaces R_j in each residual set $\mathcal{R}_{j,K}$ for the approximation class. We can then apply the theory of Section IV-C with Proposition 1 to calculate smallest allowable M via Theorem 5.

Proposition 2: The number of K-dimensional subspaces that comprise $\mathcal{R}_{j,K}$ obeys

$$R_j \leq \begin{cases} \frac{(2e)^{K(2j+1)}}{(Kj+K+1)(Kj+1)} & \text{if } 1 \leq j < \left\lfloor \frac{\log_2 N}{K} \right\rfloor, \\ \frac{2^{(3j+2)K+8} e^{jK}}{(Kj+1)K(j+1)e^2} & \text{if } j = \left\lfloor \frac{\log_2 N}{K} \right\rfloor, \\ \frac{4^{(2j+1)K+8}}{K^2 j(j+1)e^4} & \text{if } j > \left\lfloor \frac{\log_2 N}{K} \right\rfloor. \end{cases} \tag{17}$$

Using Proposition 2 and Theorem 5, we obtain the following condition for the matrix Φ to have the RAmP, which is proved in Appendix VI.

Proposition 3: Let Φ be an $M \times N$ matrix with i.i.d. subgaussian entries.. If

$$M \geq \begin{cases} \frac{2}{\left(\sqrt{1+\epsilon_K}-1\right)^2}\left(10K + 2\ln\frac{N}{K(K+1)(2K+1)} + t\right) & \text{if } K \leq \log_2 N, \\ \frac{2}{\left(\sqrt{1+\epsilon_K}-1\right)^2}\left(10K + 2\ln\frac{601N}{K^3} + t\right) & \text{if } K > \log_2 N, \end{cases}$$

then the matrix Φ has the (ϵ_K, s)-RAmP for model $\mathcal{T}$ and all $s > 0.5$ with probability $1 - e^{-t}$.

Both cases give a simplified bound on the number of measurements required as $M = \mathcal{O}(K)$, which is a substantial improvement over the $M = \mathcal{O}(K \log(N/K))$ required by conventional CS recovery methods. Thus, when Φ satisfies Proposition 3, we have the guarantee (14) for sampled Besov space signals from $B_q^s(L_p([0,1]))$.

E. Experiments

We now present the results of a number of numerical experiments that illustrate the effectiveness of a tree-based recovery algorithm. Our consistent observation is that explicit incorporation of the model in the recovery process significantly improves the quality of recovery for a given number of measurements. In addition, model-based recovery remains stable when the inputs are no longer tree-sparse, but rather are tree-compressible and/or corrupted with differing levels of noise. We employ the model-based CoSaMP recovery of Algorithm 1 with a CSSA-based approximation step in all experiments.

We first study one-dimensional signals that match the connected wavelet-tree model described above. Among such signals is the class of piecewise smooth functions, which are commonly encountered in analysis and practice.

Figure 1 illustrates the results of recovering the tree-compressible *HeaviSine* signal of length $N = 1024$ from $M = 80$ noise-free random Gaussian measurements using CoSaMP, ℓ_1-norm minimization using the `l1_eq` solver from the ℓ_1-Magic toolbox,[3] and our tree-based recovery algorithm. It is clear that the number of measurements ($M = 80$) is far fewer than the minimum number required by CoSaMP and ℓ_1-norm minimization to accurately recover the signal. In contrast, tree-based recovery using $K = 26$ is accurate and uses fewer iterations to converge

[3] *http://www.acm.caltech.edu/l1magic.*

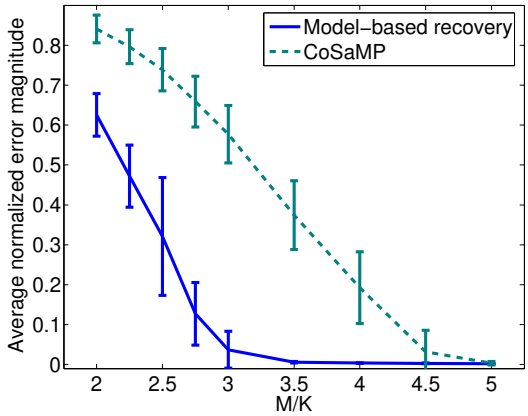


Fig. 3. *Performance of CoSaMP vs. wavelet tree-based recovery on a class of piecewise-cubic signals as a function of* M/K.

than conventional CoSaMP. Moreover, the normalized magnitude of the squared error for tree-based recovery is equal to 0.037, which is remarkably close to the error between the noise-free signal and its *best* K-term tree-approximation (0.036).

Figure 3 illustrates the results of a Monte Carlo simulation study on the impact of the number of measurements M on the performance of model-based and conventional recovery for a class of tree-sparse piecewise-polynomial signals. Each data point was obtained by measuring the normalized recovery error of 500 sample trials. Each sample trial was conducted by generating a new piecewise-polynomial signal with five polynomial pieces of cubic degree and randomly placed discontinuities, computing its best K-term tree-approximation using the CSSA, and then measuring the resulting signal using a matrix with i.i.d. Gaussian entries. Model-based recovery attains near-perfect recovery at $M = 3K$ measurements, while CoSaMP only matches this performance at $M = 5K$. We defer a full Monte Carlo comparison of our method with the much more computationally demanding ℓ_1-norm minimization to future work. In practice, we have noticed that CoSaMP and ℓ_1-norm minimization offer similar recovery trends; consequently, we can expect that model-based recovery will offer a similar degree of improvement over ℓ_1-norm minimization.

Further, we demonstrate that model-based recovery performs stably in the presence of

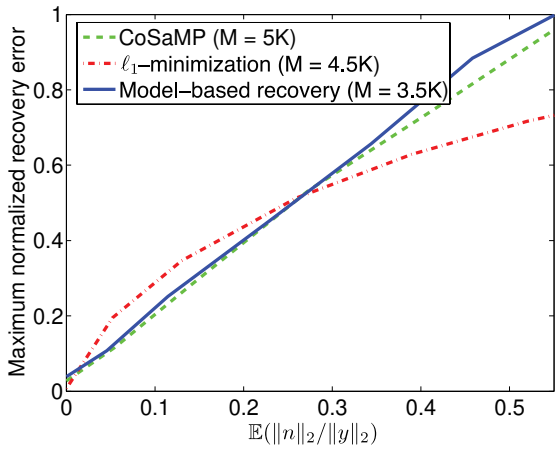


Fig. 4. *Robustness to measurement noise for standard and wavelet tree-based CS recovery algorithms. We plot the maximum normalized recovery error over 200 sample trials as a function of the expected signal-to-noise ratio. The linear growth demonstrates that model-based recovery possesses the same robustness to noise as CoSaMP and ℓ_1-norm minimization.*

measurement noise. We generated sample piecewise-polynomial signals as above, computed their best K-term tree-approximations, computed M measurements of each approximation, and finally added Gaussian noise of expected norm $\|n\|_2$ to each measurement. Then, we recovered the signal using CoSaMP and model-based recovery and measured the recovery error in each case. For comparison purposes, we also tested the recovery performance of a ℓ_1-norm minimization algorithm that accounts for the presence of noise, which has been implemented as the `l1_qc` solver in the ℓ_1-Magic toolbox. First, we determined the lowest value of M for which the respective algorithms provided near-perfect recovery in the absence of noise in the measurements. This corresponds to $M = 3.5K$ for model-based recovery, $M = 5K$ for CoSaMP, and $M = 4.5K$ for ℓ_1 minimization. Next, we generated 200 sample tree-modeled signals, computed M *noisy* measurements, recovered the signal using the given algorithm and recorded the recovery error. Figure 4 illustrates the growth in maximum normalized recovery error (over the 200 sample trials) as a function of the expected measurement signal-to-noise ratio for the tree algorithms. We observe similar stability curves for all three algorithms, while noting that model-based recovery offers this kind of stability using significantly fewer measurements.

(a) *Peppers* (b) CoSaMP (RMSE $= 22.8$) (c) model-based recovery (RMSE $= 11.1$)

Fig. 5. *Example performance of standard and model-based recovery on images. (a)* $N = 128 \times 128 = 16384$*-pixel* Peppers *test image. Image recovery from* $M = 5000$ *compressive measurements using (b) conventional CoSaMP and (c) our wavelet tree-based algorithm.*

Finally, we turn to two-dimensional images and a wavelet quadtree model. The connected wavelet-tree model has proven useful for compressing natural images [27]; thus, our algorithm provides a simple and provably efficient method for recovering a wide variety of natural images from compressive measurements. An example of recovery performance is given in Figure 5. The test image (*Peppers*) is of size $N = 128 \times 128 = 16384$ pixels, and we computed $M = 5000$ random Gaussian measurements. Model-based recovery again offers higher performance than standard signal recovery algorithms like CoSaMP, both in terms of recovery mean-squared error and visual quality.

VI. Example: Block-Sparse Signals and Signal Ensembles

In a *block-sparse* signal, the locations of the significant coefficients cluster in blocks under a specific sorting order. Block-sparse signals have been previously studied in CS applications, including DNA microarrays and magnetoencephalography [7, 8]. An equivalent problem arises in CS for signal ensembles, such as sensor networks and MIMO communication [8, 9, 34]. In this case, several signals share a common coefficient support set. For example, when a frequency-sparse acoustic signal is recorded by an array of microphones, then all of the recorded signals

contain the same Fourier frequencies but with different amplitudes and delays. Such a signal ensemble can be re-shaped as a single vector by concatenation, and then the coefficients can be rearranged so that the concatenated vector exhibits block sparsity.

It has been shown that the block-sparse structure enables signal recovery from a reduced number of CS measurements, both for the single signal case [7, 8] and the signal ensemble case [9], through the use of specially tailored recovery algorithm [7, 8, 35]. However, the robustness guarantees for such algorithms either are restricted to exactly sparse signals and noiseless measurements, do not have explicit bounds on the number of necessary measurements, or are asymptotic in nature.

In this section, we formulate the block sparsity signal model as a union of subspaces and pose an approximation algorithm on this union of subspaces. The approximation algorithm is used to implement block-based signal recovery. We also define the corresponding class of block-compressible signals and quantify the number of measurements necessary for robust recovery.

A. Block-sparse signals

Consider a class $\mathcal{S}$ of signal vectors $x \in \mathbb{R}^{JN}$, with J and N integers. This signal can be reshapped into a $J \times N$ matrix X, and we use both notations interchangeably in the sequel. We will restrict entire columns of X to be part of the support of the signal as a group. That is, signals X in a block-sparse model have entire columns as zeros or nonzeros. The measure of sparsity for X is its number of nonzero columns. More formally, we make the following definition.

Definition 11: [7, 8] Define the set of K*-block sparse signals* as

$$\mathcal{S}_K = \{X = [x_1 \ \ldots \ x_N] \in \mathbb{R}^{J\times N} \text{ such that } x_n = 0 \text{ for } n \notin \Omega, \Omega \subseteq \{1, \ldots, N\}, |\Omega| = K\}.$$

It is important to note that a K-block sparse signal has sparsity KJ, which is dependent on the size of the block J. We can extend this formulation to ensembles of J, length-N signals with common support. Denote this signal ensemble by $\{\widetilde{x}_1, \ldots, \widetilde{x}_J\}$, with $\widetilde{x}_j \in \mathbb{R}^N$, $1 \leq j \leq J$. We formulate a matrix representation $\widetilde{X}$ of the ensemble that features the signal $\widetilde{x}_j$ in its jth row: $\widetilde{X} = [\widetilde{x}_1 \ \ldots \widetilde{x}_N]^T$. The matrix $\widetilde{X}$ features the same structure as the matrix X obtained

from a block-sparse signal; thus, the matrix $\widetilde{X}$ can be converted into a block-sparse vector $\widetilde{x}$ that represents the signal ensemble.

B. Block-based approximation

To pose a the block-based approximation algorithm, we need to define the mixed norm of a matrix.

Definition 12: The (p, q) *mixed norm* of the matrix $X = [x_1 \; x_2 \; \ldots \; x_N]$ is defined as

$$\|X\|_{(p,q)} = \left(\sum_{n=1}^{N} \|x_n\|_p^q \right)^{1/q}.$$

When $q = 0$, $\|X\|_{(p,0)}$ simply counts the number of nonzero columns in X.

We immediately find that $\|X\|_{(p,p)} = \|x\|_p$, with x the vectorization of X. Intuitively, we pose the algorithm $\mathbb{S}(X, K)$ to obtain the best block-based approximation of the signal X as follows:

$$X_K^{\mathcal{S}} = \arg \min_{\bar{X} \in \mathbb{R}^{J \times N}} \|X - \bar{X}\|_{(2,2)} \text{ subject to } \|\bar{X}\|_{(2,0)} \leq K. \tag{18}$$

It is easy to show that to obtain the approximation, it suffices to perform column-wise hard thresholding: let ρ be the Kth largest ℓ_2-norm among the columns of X. Then our approximation algorithm is $\mathbb{S}(X, K) = X_K^{\mathcal{S}} = [x_{K,1}^{\mathcal{S}} \; \ldots x_{K,N}^{\mathcal{S}}]$, where

$$x_{K,n}^{\mathcal{S}} = \begin{cases} x_n & \|x_n\|_2 \geq \rho, \\ 0 & \|x_n\|_2 < \rho, \end{cases}$$

for each $1 \leq j \leq J$ and $1 \leq n \leq N$. Alternatively, a recursive approximation algorithm can be obtained by sorting the columns of X by their ℓ_2 norms, and then selecting the largest columns. The complexity of this sorting process is $\mathcal{O}(NJ + N \log N)$.

C. Block-compressible signals

The approximation class under the block-compressible model corresponds to signals with blocks whose ℓ_2 norm has a power-law decay rate.

Definition 13: We define the set of s-block compressible signals as

$$\mathfrak{S}_s = \{X = [x_1 \ \ldots \ x_N] \in \mathbb{R}^{J\times N} \text{ s.t. } \|x_{\mathcal{I}(i)}\|_2 \leq S i^{-s-1/2}, 1 \leq i \leq N, S < \infty\},$$

where $\mathcal{I}$ indexes the sorted column norms.

We say that X is an s-block compressible signal if $X \in \mathfrak{S}_s$. For such signals, we have $\|X - X_K\|_{(2,2)} = \sigma_{\mathcal{S}_K}(x) \leq S_1 K^{-s}$, and $\|X - X_K\|_{(2,1)} \leq S_2 K^{1/2-s}$. Note that the block-compressible signal model does not impart a structure to the decay of the signal coefficients, so that the sets $\mathcal{R}_{j,K}$ are equal for all values of j; due to this property, the $(\delta_{\mathcal{S}_K}, s)$-RAmP is implied by the $\mathcal{S}_K$-RIP. Taking this into account, we can derive the following result from [10], which is proven similarly to Theorem 4.

Theorem 6: Let x be a signal from model $\mathcal{S}$, and let $y = \Phi x + n$ be a set of noisy CS measurements. If Φ has the $\mathcal{S}_K^4$-RIP with $\delta_{\mathcal{S}_K^4} \leq 0.1$, then the estimate obtained from iteration i of block-based CoSaMP, using the approximation algorithm (18), satisfies

$$\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 20\left(\|X - X_K^{\mathcal{S}}\|_{(2,2)} + \frac{1}{\sqrt{K}}\|X - X_K^{\mathcal{S}}\|_{(2,1)} + \|n\|_2\right).$$

Thus, the algorithm provides a recovered signal of similar quality to approximations of X by a small number of nonzero columns. When the signal x is K-block sparse, we have that $||X - X_K^{\mathcal{S}}\|_{(2,2)} = ||X - X_K^{\mathcal{S}}\|_{(2,1)} = 0$, obtaining the same result as Theorem 4, save for a constant factor.

D. Stable block-based recovery from compressive measurements

Since Theorem 6 poses the same requirement on the measurement matrix Φ for sparse and compressible signals, the same number of measurements M is required to provide performance guarantees for block-sparse and block-compressible signals. The class $\mathcal{S}_K$ contains $S = \binom{N}{K}$ subspaces of dimension JK. Thus, a subgaussian random matrix has the $\mathcal{S}_K$-RIP property with constant $\delta_{\mathcal{S}_K}$ and probability $1 - e^{-t}$ if the number of measurements obeys

$$M \geq \frac{2}{c\delta_{\mathcal{S}_K}^2}\left(K\left(\ln\frac{2N}{K} + J\ln\frac{12}{\delta_{\mathcal{S}_K}}\right) + t\right). \tag{19}$$

The first term in this bound matches the order of the bound for conventional CS, while the second term introduces a linear dependence on the size of the block J. This shows that the number of measurements required for robust recovery scales as $M = \mathcal{O}\left(KJ + K\log(N/K)\right)$, which is a substantial improvement over the $M = \mathcal{O}\left(JK\log(N/K)\right)$ that would be required by conventional CS recovery methods. When the size of the block J is larger than $\log(N/K)$, then this term becomes $\mathcal{O}\left(KJ\right)$; that is, it is linear on the total sparsity of the block-sparse signal.

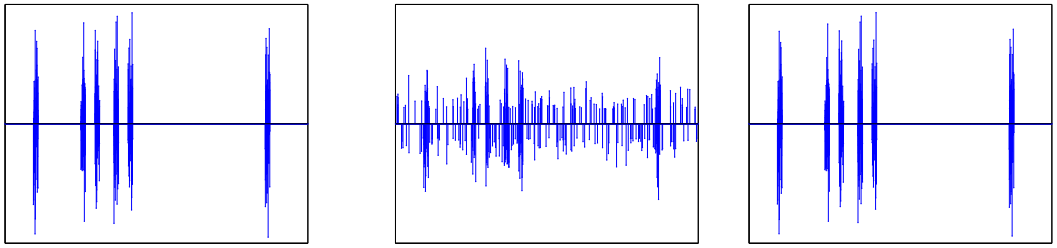
We note in passing that the bound on the number of measurements (19) assumes a dense subgaussian measurement matrix, while the measurement matrices used in [9] have a block-diagonal. structure. To obtain measurements from an $M \times JN$ dense matrix in a distributed setting, it suffices to partition the matrix into J pieces of size $M \times N$ and calculate the CS measurements at each sensor with a corresponding matrix; these individual measurements are then summed to obtain the complete measurement vector. For large J, (19) implies that the total number of measurements required for recovery of the signal ensemble is lower than the bound for the case where each signal recovery is performed independently for each signal (M = $\mathcal{O}\left(JK\log(N/K)\right)$).

E. Experiments

We conducted several numerical experiments comparing model-based recovery to CoSaMP in the context of block-sparse signals. We employ the model-based CoSaMP recovery of Algorithm 1 with the block-based approximation algorithm (18) in all cases. For brevity, we exclude a thorough comparison of our model-based algorithm with ℓ_1-based optimization and defer it to future work. In practice, we observed that our algorithm performs several times faster than convex optimization-based procedures.

Figure 6 illustrates an $N = 4096$ signal that exhibits block sparsity, and its recovered version using CoSaMP and model-based recovery. The block size $J = 64$ and there were $K = 6$ active blocks in the signal. We observe the clear advantage of using the block-sparsity model in signal recovery.

We now consider block-compressible signals. An example recovery is illustrated in Figure 7.



(a) original block-sparse signal (b) CoSaMP (RMSE $= 0.723$) (c) model-based recovery (RMSE $= 0.015$)

Fig. 6. *Example performance of model-based signal recovery for a block-sparse signal. (a) Example block-compressible signal of length $N = 4096$ with $K = 6$ nonzero blocks of size $J = 64$. Recovered signal from $M = 960$ measurements using (b) conventional CoSaMP recovery and (c) block-based recovery.*

In this case, the ℓ_2-norms of the blocks decay according to a power law, as described above. Again, the number of measurements is far below the minimum number required to guarantee stable recovery through conventional CS recovery. However, enforcing the model in the approximation process results in a solution that is very close to the best 5-block approximation of the signal.

Figure 8 indicates the decay in recovery error as a function of the numbers of measurements for CoSaMP and model-based recovery. We generated sample block-sparse signals as follows: we randomly selected a set of K blocks, each of size J, and endow them with coefficients that follow an i.i.d. Gaussian distribution. Each sample point in the curves is generated by performing 200 trials of the corresponding algorithm. As in the connected wavelet-tree case, we observe clear gains using model-based recovery, particularly for low-measurement regimes; CoSaMP matches model-based recovery only for $M \geq 5K$.

VII. Conclusions

In this paper, we have aimed to demonstrate that there are significant performance gains to be made by exploiting more realistic and richer signal models beyond the simplistic sparse and compressible models that dominate the CS literature. Building on the unions of subspaces results of [5] and the proof machinery of [10], we have taken some of the first steps towards

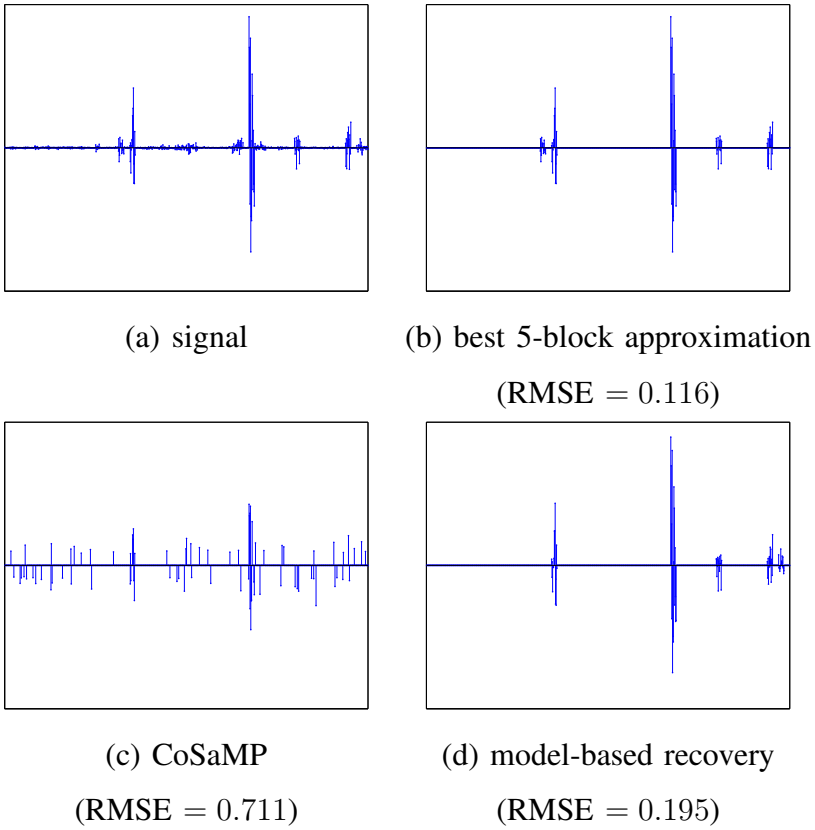


Fig. 7. *Example performance of model-based signal recovery for block-compressible signals. (a) Example block-compressible signal, length $N = 1024$. (b) Best block-based approximation with $K = 5$ blocks. Recovered signal from $M = 200$ measurements using both (c) conventional CoSaMP recovery and (d) block-based recovery.*

what promises to be a general theory for model-based CS by introducing the notion of a model-compressible signal and the associated restricted amplification property (RAmP) condition it imposes on the measurement matrix Φ.

For the volumes of natural and manmade signals and images that are wavelet-sparse or compressible, our tree-based CoSaMP and IHT algorithms offer performance that significantly exceeds today's state-of-the-art while requiring only $M = \mathcal{O}(K)$ rather than $M = \mathcal{O}(K \log(N/K))$ random measurements. For block-sparse signals and signal ensembles, our block-based CoSaMP and IHT algorithms offer not only excellent performance but also require just $M = \mathcal{O}(JK)$ measurements, where JK is the signal sparsity. Furthermore, block-based recovery can recovery signal ensembles using fewer measurements than the number required

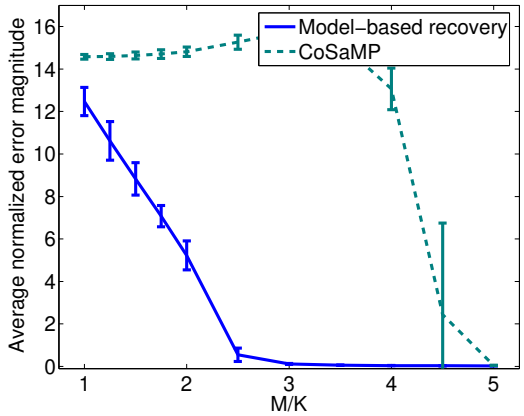


Fig. 8. *Performance of CoSaMP and block-based recovery on a class of block-sparse signals as a function of M/K. Standard CS recovery does not match the performance of block-based recovery until $M = 5K$.*

when each signal is recovered independently.

There are many avenues for future work on model-based CS. We have only considered the recovery of signals from models that can be geometrically described as a union of subspaces; possible extensions include other, more complex geometries (for example, high-dimensional polytopes, nonlinear manifolds.) We also expect that the core of our proposed algorithms — a model-enforcing approximation step — can be integrated into other iterative algorithms, such as relaxed ℓ_1-norm minimization methods. Furthermore, our framework will benefit from the formulation of new signal models that are endowed with efficient model-based approximation algorithms.

APPENDIX I
PROOF OF THEOREM 2

To prove this theorem, we will study the distribution of the maximum singular value of a submatrix Φ_T of a matrix with i.i.d. Gaussian entries Φ corresponding to the columns indexed by T. From this we obtain the probability that RAmP does not hold for a fixed support T. We will then evaluate the same probability for all supports T of elements of $\mathcal{R}_{j,K}$, where the desired bound on the amplification is dependent on the value of j. This gives us the probability that the

RAmP does not hold for a given residual subspace set $\mathcal{R}_{j,K}$. We fix the probability of failure on each of these sets; we then obtain probability that the matrix Φ does not have the RAmP using a union bound. We end by obtaining conditions on the number of rows M of Φ to obtain a desired probability of failure.

We begin from the following concentration of measure for the largest singular value of a $M \times K$ submatrix Φ_T, $|T| = K$, of an $M \times N$ matrix Φ with i.i.d. subgaussian entries that are properly normalized [18, 36, 37]:

$$P\left(\sigma_{\max}(\Phi_T) > 1 + \sqrt{\frac{K}{M}} + \tau + \beta\right) \leq e^{-M\tau^2/2}.$$

For large enough M, $\beta \ll 1$; thus we ignore this small constant in the sequel. By letting $\tau = j^r\sqrt{1+\epsilon_K} - 1 - \sqrt{\frac{K}{M}}$ (with the appropriate value of j for T), we obtain

$$P\left(\sigma_{\max}(\Phi_T) > j^r\sqrt{1+\epsilon_K}\right) \leq e^{-\frac{M}{2}\left(j^r\sqrt{1+\epsilon_K}-1-\sqrt{\frac{K}{M}}\right)^2}.$$

We use a union bound over all possible R_j supports for $u \in \mathcal{R}_{j,K}$ to obtain the probability that Φ does not amplify the norm of u by more than $j^r\sqrt{1+\epsilon_K}$:

$$P\left(\|\Phi u\|_2 > \left(j^r\sqrt{1+\epsilon_K}\right)\|u\|_2 \ \forall\ u \in \mathcal{R}_{j,K}\right) \leq R_j e^{-\frac{1}{2}\left(\sqrt{M}(j^r\sqrt{1+\epsilon_K}-1)-\sqrt{K}\right)^2}.$$

Bound the right hand side by a constant μ; this requires

$$R_j \leq e^{\frac{1}{2}\left(\sqrt{M}(j^r\sqrt{1+\epsilon_K}-1)-\sqrt{K}\right)^2}\mu \tag{20}$$

for each j. We use another union bound among the residual subspaces $\mathcal{R}_{j.K}$ to measure the probability that the RAmP does not hold:

$$P\left(\|\Phi u\|_2 > \left(j^r\sqrt{1+\epsilon_K}\right)\|u\|_2 \ \forall\ u \in \mathcal{R}_{j,K},\ \forall\ j, 1 \leq j \leq \lceil N/K \rceil\right) \leq \left\lceil \frac{N}{K} \right\rceil \mu.$$

To bound this probability by e^{-t}, we need $\mu = \frac{K}{N}e^{-t}$; plugging this into (20), we obtain

$$R_j \leq e^{\frac{1}{2}\left(\sqrt{M}(j^r\sqrt{1+\epsilon_K}-1)-\sqrt{K}\right)^2}\frac{K}{N}e^{-t}$$

for each j. Simplifying, we obtain that for Φ to posess the RAmP with probability $1 - e^{-t}$, the following must hold for all j:

$$M \geq \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(\sqrt{2\left(\ln\frac{R_jN}{K}+t\right)}+\sqrt{K}\right)^2. \tag{21}$$

Since $(\sqrt{a}+\sqrt{b})^2 \leq 2a+2b$ for $a,b>0$, then the hypothesis (12) implies (21), proving the theorem. $\square$

APPENDIX II
PROOF OF THEOREM 5

In this proof, we denote $\mathbb{M}(x,K)=x_K$ for brevity. To bound $\|\Phi(x-x_K)\|_2$, we write x as

$$x = x_K + \sum_{j=2}^{\lceil N/K \rceil} x_{T_j},$$

where

$$x_{T_j} = x_{jK} - x_{(j-1)K}, j=2,\ldots,\lceil N/K \rceil$$

is the difference between the best jK model approximation and the best $(j-1)K$ model approximation. Additionally, each piece $x_{T_j} \in \mathcal{R}_{j,K}$. Therefore, since Φ satisifes the $(\epsilon_K, s-1)$ RAmP, we obtain

$$\|\Phi(x-x_K)\|_2 = \left\| \Phi\left(\sum_{j=2}^{\lceil N/K \rceil} x_{T_j} \right) \right\|_2 \leq \sum_{j=2}^{\lceil N/K \rceil} \|\Phi x_{T_j}\|_2 \leq \sum_{j=2}^{\lceil N/K \rceil} \sqrt{1+\epsilon_K} j^{s-1} \|x_{T_j}\|_2. \quad (22)$$

Since $x \in \mathfrak{M}_s$, the norm of each piece can be bounded as

$$\|x_{T_j}\|_2 = \|x_{jK} - x_{(j-1)K}\|_2 \leq \|x - x_{(j-1)K}\|_2 + \|x - x_{jK}\|_2 \leq |x|_{\mathfrak{M}_s} K^{-s} \left((j-1)^{-s} + j^{-s} \right).$$

Applying this bound in (22), we obtain

$$\begin{aligned} \|\Phi(x-x_K)\|_2 &\leq \sqrt{1+\epsilon_K} \sum_{j=2}^{\lceil N/K \rceil} j^{s-1} \|x_{T_j}\|_2, \\ &\leq \sqrt{1+\epsilon_K} |x|_{\mathfrak{M}_s} K^{-s} \sum_{j=2}^{\lceil N/K \rceil} j^{s-1} ((j-1)^{-s} + j^{-s}), \\ &\leq \sqrt{1+\epsilon_K} |x|_{\mathfrak{M}_s} K^{-s} \sum_{j=2}^{\lceil N/K \rceil} j^{-1}. \end{aligned}$$

It is easy to show, using Euler-Maclaurin summations, that $\sum_{j=2}^{\lceil N/K \rceil} j^{-1} \leq \ln \lceil N/K \rceil$; we then obtain

$$\|\Phi(x-x_K)\|_2 \leq \sqrt{1+\epsilon_K} K^{-s} \ln \left\lceil \frac{N}{K} \right\rceil |x|_{\mathfrak{M}_s},$$

which proves the theorem. $\square$

Algorithm 2 Model-Based Iterative Hard Thresholding

Inputs: CS Matrix Φ, measurements y, model $\mathcal{M}_K$

Outpurs: K-sparse approximation $\widehat{x}$

initialize: $\widehat{x}_0 = 0$, $r = y$, $i = 0$.

while halting criterion false **do**

$i \leftarrow i + 1$

$b \leftarrow \widehat{x}_{i-1} + \Phi^T r$ {form signal estimate}

$\widehat{x}_i \leftarrow \mathbb{M}(b, K)$ {prune signal estimate according to signal model}

$r \leftarrow y - \Phi\widehat{x}_i$ {update measurement residual}

end while

return $\widehat{x} \leftarrow \widehat{x}_i$

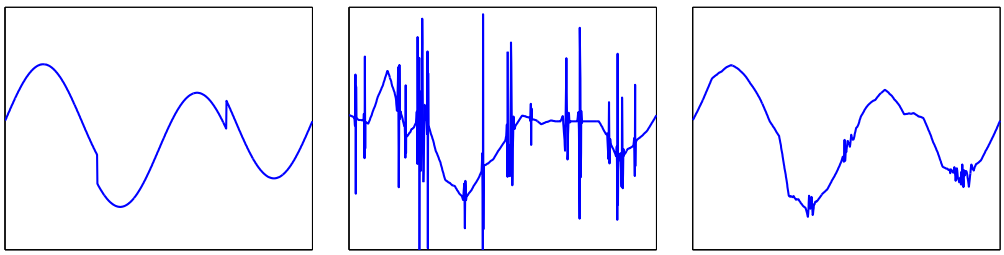
APPENDIX III
MODEL-BASED ITERATIVE HARD THRESHOLDING

Our proposed model-based iterative hard thresholding (IHT) is given in Algorithm 2. For this algorithm, Theorems 4, 5, and 6 can be proven with only a few modifications: Φ must have the $\mathcal{M}_K^3$-RIP with $\delta_{\mathcal{M}_K^3} \leq 0.1$, and the constant factor in the bound changes from 15 to 4 in Theorem 4, from 35 to 10 in Theorem 5, and from 20 to 5 in Theorem 6.

To illustrate the performance of the algorithm, we repeat the *HeaviSine* experiment from Figure 1. Recall that $N = 1024$, and $M = 80$ for this example. The advantages of using our tree-model-based approximation step (instead of mere hard thresholding) are evident from Figure 9. In practice, we have observed that our model-based algorithm converges in fewer steps than IHT and yields much more accurate results in terms of recovery error.

APPENDIX IV
PROOF OF THEOREM 4

The proof of this theorem is identical to that of the CoSaMP algorithm in [10, Section 4.6], and requires a set of six lemmas. The sequence of Lemmas 1–6 below are modifications of



(a) original (b) IHT (RMSE $= 0.627$) (c) model-based IHT (RMSE $= 0.080$)

Fig. 9. *Example performance of model-based IHT. (a) Piecewise-smooth* HeaviSine *test signal, length* $N = 1024$*. Signal recovered from* $M = 80$ *measurements using both (b) standard and (c) model-based IHT recovery. Root mean-squared error (RMSE) values are normalized with respect to the* ℓ_2 *norm of the signal.*

the lemmas in [10] that are restricted to the signal model. Lemma 4 does not need any changes from [10], so we state it without proof. The proof of Lemmas 3–6 use the properties in Lemmas 1 and 2, which are simple to prove.

Lemma 1: Suppose Φ has $\mathcal{M}$-RIP with constant $\delta_{\mathcal{M}}$. Let Ω be a support corresponding to a subspace in $\mathcal{M}$. Then we have the following handy bounds.

$$\begin{aligned}
\|\Phi_\Omega^T u\|_2 &\leq \sqrt{1+\delta_{\mathcal{M}}}\|u\|_2, \\
\|\Phi_\Omega^\dagger u\|_2 &\leq \frac{1}{\sqrt{1-\delta_{\mathcal{M}}}}\|u\|_2, \\
\|\Phi_\Omega^T \Phi_\Omega u\|_2 &\leq (1+\delta_{\mathcal{M}})\|u\|_2, \\
\|\Phi_\Omega^T \Phi_\Omega u\|_2 &\geq (1-\delta_{\mathcal{M}})\|u\|_2, \\
\|(\Phi_\Omega^T \Phi_\Omega)^{-1} u\|_2 &\leq \frac{1}{1+\delta_{\mathcal{M}}}\|u\|_2, \\
\|(\Phi_\Omega^T \Phi_\Omega)^{-1} u\|_2 &\geq \frac{1}{1-\delta_{\mathcal{M}}}\|u\|_2.
\end{aligned}$$

Lemma 2: Suppose Φ has $\mathcal{M}_K^2$-RIP with constant $\delta_{\mathcal{M}_K^2}$. Let Ω be a support corresponding to a subspace in $\mathcal{M}_K$, and let $x \in \mathcal{M}_K$. Then $\|\Phi_\Omega^T \Phi x|_{\Omega^C}\|_2 \leq \delta_{\mathcal{M}_K^2}\|x|_{\Omega^C}\|_2$.

We begin the proof of Theorem 4 by fixing an iteration $i \geq 1$ of model-based CoSaMP. We write $\widehat{x} = \widehat{x}_{i-1}$ for the signal estimate at the beginning of the ith iteration. Define the signal residual $s = x - \widehat{x}$, which implies that $s \in \mathcal{M}_K^2$. We note that we can write $r = y - \Phi\widehat{x} = \Phi(x - \widehat{x}) + n = \Phi s + n$.

Lemma 3: (Identification) The set $\Omega = \text{supp}(\mathbb{M}_2(e, K))$, where $e = \Phi^T r$, identifies a subspace in $\mathcal{M}_K^2$, and obeys

$$\|s|_{\Omega^C}\|_2 \leq 0.2223\|s\|_2 + 2.34\|n\|_2.$$

Proof of Lemma 3: Define the set $\Pi = \text{supp}(s)$. Let $e_\Omega = \mathbb{M}_2(e, K)$ be the model-based approximation to e with support Ω, and similarly let e_Π be the approximation to e with support Π. Each approximation is equal to e for the coefficients in the support, and zero elsewhere. Since Ω is the support of the best approximation in $\mathcal{M}_K^2$, we must have:

$$\begin{aligned}
\|e - e_\Omega\|_2^2 &\leq \|e - e_\Pi\|_2^2, \\
\sum_{n=1}^{N}(e[n] - e_\Omega[n])^2 &\leq \sum_{n=1}^{N}(e[n] - e_\Pi[n])^2, \\
\sum_{n \notin \Omega} e[n]^2 &\leq \sum_{n \notin \Pi} e[n]^2, \\
\sum_{n=1}^{N} e[n]^2 - \sum_{n \notin \Omega} e[n]^2 &\geq \sum_{n=1}^{N} e[n]^2 - \sum_{n \notin \Pi} e[n]^2, \\
\sum_{n \in \Omega} e[n]^2 &\geq \sum_{n \in \Pi} e[n]^2, \\
\sum_{n \in \Omega} e[n]^2 &\geq \sum_{n \in \Pi} e[n]^2, \\
\sum_{n \in \Omega \setminus \Pi} e[n]^2 &\geq \sum_{n \in \Pi \setminus \Omega} e[n]^2, \\
\|e|_{\Omega \setminus \Pi}\|_2^2 &\geq \|e|_{\Pi \setminus \Omega}\|_2^2,
\end{aligned}$$

where $\Omega \setminus \Pi$ denotes the set difference of Ω and Π. These signals are in $\mathcal{M}_K^4$ (since they arise as the difference of two elements from $\mathcal{M}_K^2$); therefore, we can apply the $\mathcal{M}_K^4$-RIP constants

and Lemmas 1 and 2 to provide the following bounds on both sides (see [10] for details):

$$\|e|_{\Omega\setminus\Pi}\|_2 \;\leq\; \delta_{\mathcal{M}_K^4}\|s\|_2 + \sqrt{1+\delta_{\mathcal{M}_K^2}}\|e\|_2, \tag{23}$$

$$\|e|_{\Pi\setminus\Omega}\|_2 \;\geq\; (1-\delta_{\mathcal{M}_K^2})\|s|_{\Omega^C}\|_2 - \delta_{\mathcal{M}_K^2}\|s\|_2 - \sqrt{1+\delta_{\mathcal{M}_K^2}}\|e\|_2. \tag{24}$$

Combining (23) and (24), we obtain

$$\|s|_{\Omega^C}\|_2 \leq \frac{(\delta_{\mathcal{M}_K^2}+\delta_{\mathcal{M}_K^4})\|s\|_2 + 2\sqrt{1+\delta_{\mathcal{M}_K^2}}\|e\|_2}{1-\delta_{\mathcal{M}_K^2}}.$$

The argument is completed by noting that $\delta_{\mathcal{M}_K^2} \leq \delta_{\mathcal{M}_K^4} \leq 0.1$. □

Lemma 4: (Support Merger) Let Ω be a set of at most $2K$ indices. Then the set $\Lambda = \Omega \cup \mathrm{supp}(\widehat{x})$ contains at most $3K$ indices, and $\|x|_{\Lambda^C}\|_2 \leq \|s|_{\Omega^C}\|_2$.

Lemma 5: (Estimation) Let Λ be a support corresponding to a subspace in $\mathcal{M}_K^3$, and define the least squares signal estimate b by $b|_T = \Phi_T^\dagger y$, $b|_{T^C} = 0$. Then

$$\|x-b\|_2 \leq 1.112\|x|_{\Lambda^C}\|_2 + 1.06\|n\|_2.$$

Proof of Lemma 5: It can be shown [10] that

$$\|x-b\|_2 \leq \|x|_{\Lambda^C}\|_2 + \|(\Phi_\Lambda^T\Phi_\Lambda)^{-1}\Phi_\Lambda^T\Phi x|_{\Pi^C}\|_2 + \|\Phi_\Pi^\dagger n\|_2.$$

Since Λ is a support corresponding to a subspace in $\mathcal{M}_K^3$ and $x \in \mathcal{M}_K$, we use Lemmas 1 and 2 to obtain

$$\begin{aligned}\|x-b\|_2 \;&\leq\; \|x|_{\Lambda^C}\|_2 + \frac{1}{1-\delta_{\mathcal{M}_K^3}}\|\Phi_\Lambda^T\Phi x|_{\Pi^C}\|_2 + \frac{1}{\sqrt{1-\delta_{\mathcal{M}_K^3}}}\|n\|_2,\\ &\leq\; \left(1+\frac{\delta_{\mathcal{M}_K^4}}{1-\delta_{\mathcal{M}_K^3}}\right)\|x|_{\Pi^C}\|_2 + \frac{1}{\sqrt{1-\delta_{\mathcal{M}_K^3}}}\|n\|_2.\end{aligned}$$

Finally, note that $\delta_{\mathcal{M}_K^3} \leq \delta_{\mathcal{M}_K^4} \leq 0.1$. □

Lemma 6: (Pruning) The pruned approximation $\widehat{x}_i = \mathbb{M}(b, K)$ is such that

$$\|x-\widehat{x}_i\|_2 \leq 2\|x-b\|_2.$$

Proof of Lemma 6: Since $\widehat{x}_i$ is the best approximation in $\mathcal{M}_K$ to b, and $x \in \mathcal{M}_K$, we obtain

$$\|x - \widehat{x}_i\|_2 \leq \|x - b\|_2 + \|b - \widehat{x}_i\|_2 \leq 2\|x - b\|_2.$$

□

We use these lemmas in reverse sequence for the inequalities below:

$$\begin{aligned}
\|x - \widehat{x}_i\|_2 &\leq 2\|x - b\|_2, \\
&\leq 2(1.112\|x|_{\Lambda^C}\|_2 + 1.06\|n\|_2), \\
&\leq 2.224\|s|_{\Omega^C}\|_2 + 2.12\|n\|_2, \\
&\leq 2.224(0.2223\|s\|_2 + 2.34\|n\|_2) + 2.12\|n\|_2, \\
&\leq 0.5\|s\|_2 + 7.5\|n\|_2, \\
&\leq 0.5\|x - \widehat{x}_{i-1}\|_2 + 7.5\|n\|_2.
\end{aligned}$$

From the recursion on $\widehat{x}_i$, we obtain $\|x - \widehat{x}_i\|_2 \leq 2^{-i}\|x\|_2 + 15\|n\|_2$. This completes the proof of Theorem 4. □

APPENDIX V
PROOF OF PROPOSITION 1

When $K < \log_2 N$, the number of subtrees of size K of a binary tree of size N is the Catalan number [38]

$$T_{K,N} = \frac{1}{K+1}\binom{2K}{K} \leq \frac{(2e)^K}{K+1},$$

using Stirling's approximation. When $K > \log_2 N$, we partition this count of subtrees into the numbers of subtrees $t_{K,h}$ of size K and height h, to obtain

$$T_{K,N} = \sum_{h=\lfloor \log_2 K \rfloor + 1}^{\log_2 N} t_{K,h}$$

We obtain the following asymptotic identity from [38, page 51]:

$$\begin{aligned} t_{K,h} &= \frac{4^{K+1.5}}{h^4}\sum_{m\geq 1}\left[\frac{2K}{h^2}(2\pi m)^4 - 3(2\pi m)^2\right]e^{-\frac{K(2\pi m)^2}{h^2}} + 4^K\mathcal{O}\left(e^{-\ln^2 h}\right) \\ &\quad +4^K\mathcal{O}\left(\frac{\ln^8 h}{h^5}\right) + 4^K\mathcal{O}\left(\frac{\ln^8 h}{h^4}\right), \\ &\leq \frac{4^{K+2}}{h^4}\sum_{m\geq 1}\left[\frac{2K}{h^2}(2\pi m)^4 - 3(2\pi m)^2\right]e^{-\frac{K(2\pi m)^2}{h^2}}. \end{aligned} \tag{25}$$

We now simplify the formula slightly: we seek a bound for the sum term (which we denote by β_h for brevity):

$$\beta_h = \sum_{m\geq 1}\left[\frac{2K}{h^2}(2\pi m)^4 - 3(2\pi m)^2\right]e^{-\frac{K(2\pi m)^2}{h^2}} \leq \sum_{m\geq 1}\frac{2K}{h^2}(2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}}. \tag{26}$$

Let $m_{\max} = \frac{h}{\pi\sqrt{2K}}$, the value of m for which the term inside the sum (26) is maximum; this is not necessarily an integer. Then,

$$\begin{aligned} \beta_h &\leq \sum_{m=1}^{\lfloor m_{\max}\rfloor - 1}\frac{2K}{h^2}(2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} + \sum_{m=\lfloor m_{\max}\rfloor}^{\lceil m_{\max}\rceil}\frac{2K}{h^2}(2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + \sum_{m\geq\lceil m_{\max}\rceil+1}\frac{2K}{h^2}(2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}}, \\ &\leq \int_1^{\lfloor m_{\max}\rfloor}\frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx + \sum_{m=\lfloor m_{\max}\rfloor}^{\lceil m_{\max}\rceil}\frac{2K}{h^2}(2\pi m)^4 e^{-\frac{K(2\pi m)^2}{h^2}} \\ &\quad + \int_{\lceil m_{\max}\rceil}^{\infty}\frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx, \end{aligned}$$

where the second inequality comes from the fact that the series in the sum is strictly increasing for $m \leq \lfloor m_{\max}\rfloor$ and strictly decreasing for $m > \lceil m_{\max}\rceil$. One of the terms in the sum can be added to one of the integrals. If we have that

$$(2\pi\lfloor m_{\max}\rfloor)^4 e^{-\frac{K(2\pi\lfloor m_{\max}\rfloor)^2}{h^2}} < (2\pi\lceil m_{\max}\rceil)^4 e^{-\frac{K(2\pi(\lceil m_{\max}\rceil))^2}{h^2}}, \tag{27}$$

then we can obtain

$$\begin{aligned} \beta_h &\leq \int_1^{\lceil m_{\max}\rceil}\frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx + \frac{2K}{h^2}(2\pi\lceil m_{\max}\rceil)^4 e^{-\frac{K(2\pi\lceil m_{\max}\rceil)^2}{h^2}} \\ &\quad + \int_{\lceil m_{\max}\rceil}^{\infty}\frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx. \end{aligned}$$

When the opposite of (27) is true, we have that

$$\begin{aligned}\beta_h &\leq \int_1^{\lfloor m_{\max}\rfloor} \frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx + \frac{2K}{h^2}(2\pi\lfloor m_{\max}\rfloor)^4 e^{-\frac{K(2\pi\lfloor m_{\max}\rfloor)^2}{h^2}} \\ &\quad + \int_{\lfloor m_{\max}\rfloor}^{\infty} \frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx.\end{aligned}$$

Since the term in the sum reaches its maximum for $m_{\max}$, we will have in all three cases that

$$\beta_h \leq \int_1^{\infty} \frac{2K}{h^2}(2\pi x)^4 e^{-\frac{K(2\pi x)^2}{h^2}}dx + \frac{8h^2}{Ke^2}.$$

We perform a change of variables $u = 2\pi x$ and define $\sigma = h/\sqrt{2K}$ to obtain

$$\beta_h \leq \frac{1}{2\pi}\int_0^{\infty} \frac{1}{\sigma^2}u^4 e^{-u^2/2\sigma^2}dx + \frac{8h^2}{Ke^2} \leq \frac{1}{2\sigma\sqrt{2\pi}}\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}u^4 e^{-u^2/2\sigma^2}dx + \frac{8h^2}{Ke^2}.$$

Using the formula for the fourth central moment of a Gaussian distribution:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma}u^4 e^{-u^2/2\sigma^2}dx = 3\sigma^4,$$

we obtain

$$\beta_h \leq \frac{3\sigma^3}{2\sqrt{2\pi}} + \frac{8h^2}{Ke^2} = \frac{3h^3}{8\sqrt{\pi K^3}} + \frac{8h^2}{Ke^2}.$$

Thus, (25) simplifies to

$$t_{K,h} \leq \frac{4^K}{K}\left(\frac{6}{h\sqrt{\pi K}} + \frac{128}{h^2e^2}\right).$$

Correspondingly, $T_{K,N}$ becomes

$$\begin{aligned}T_{K,N} &\leq \sum_{h=\lfloor\log_2 K\rfloor+1}^{\log_2 N} \frac{4^K}{K}\left(\frac{6}{h\sqrt{\pi K}} + \frac{128}{h^2e^2}\right), \\ &\leq \frac{4^K}{K}\left(\frac{6}{\sqrt{\pi K}}\sum_{h=\lfloor\log_2 K\rfloor+1}^{\log_2 N}\frac{1}{h} + \frac{128}{e^2}\sum_{h=\lfloor\log_2 K\rfloor+1}^{\log_2 N}\frac{128}{h^2e^2}\right).\end{aligned}$$

It is easy to show, using Euler-Maclaurin summations, that

$$\sum_{j=a}^{b} j^{-1} \leq \ln\frac{b}{a-1} \text{ and } \sum_{j=a}^{b} j^{-1} \leq \frac{1}{a-1};$$

we then obtain

$$T_{K,N} \leq \frac{4^K}{K}\left(\frac{6}{\sqrt{\pi K}}\ln\frac{\log_2 N}{\lfloor\log_2 K\rfloor} + \frac{128}{e^2\lfloor\log_2 K\rfloor}\right) \leq \frac{4^{K+4}}{Ke^2\lfloor\log_2 K\rfloor} \leq \frac{4^{K+4}}{Ke^2}.$$

This proves the proposition. □

APPENDIX VI

PROOF OF PROPOSITION 3

We wish to find the value of the bound (12) for the subspace count given in (17). We obtain $M \geq \max_{1\leq j\leq \lceil N/K\rceil} M_j$, where M_j follows one of these three regimes:

$$M_j = \begin{cases} \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(2K + 4\ln\frac{(2e)^{K(2j+1)}N}{K(Kj+1)(Kj+K+1)} + 2t\right) & \text{if } j < \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(2K + 4\ln\frac{2^{(3j+2)K+8}e^{jK}N}{K(Kj+1)(Kj+K+1)e^2} + 2t\right) & \text{if } j = \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(2K + 4\ln\frac{4^{(2j+1)K+8}N}{K^3j(j+1)e^4} + 2t\right) & \text{if } j > \left\lfloor \frac{\log_2 N}{K}\right\rfloor. \end{cases}$$

We separate the terms that are linear on K and j, and obtain

$$\begin{aligned} M_j &= \begin{cases} \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(K(3+4\ln 2) + 8Kj(1+\ln 2) + 4\ln\frac{N}{K(Kj+1)(Kj+K+1)} + 2t\right) & \text{if } j < \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(2K(1+4\ln 2) + 4Kj(1+\ln 8) + 4\ln\frac{256N}{K(Kj+1)(Kj+K+1)e^2} + 2t\right) & \text{if } j = \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^r\sqrt{1+\epsilon_K}-1\right)^2}\left(2K(1+4\ln 2) + 16Kj\ln 2 + 4\ln\frac{65536N}{K^3j(j+1)e^4} + 2t\right) & \text{if } j > \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \end{cases} \\ &= \begin{cases} \frac{1}{\left(j^{s-0.5}\sqrt{1+\epsilon_K}-j^{-0.5}\right)^2}\left(8K(1+\ln 2) + \frac{K(3+4\ln 2)}{j} + \frac{4}{j}\ln\frac{N}{K(Kj+1)(Kj+K+1)} + \frac{2t}{j}\right) & \text{if } j < \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^{s-0.5}\sqrt{1+\epsilon_K}-j^{-0.5}\right)^2}\left(4K(1+\ln 8) + \frac{2K(1+4\ln 2)}{j} + \frac{4}{j}\ln\frac{256N}{K(Kj+1)(Kj+K+1)e^2} + \frac{2t}{j}\right) & \text{if } j = \left\lfloor \frac{\log_2 N}{K}\right\rfloor, \\ \frac{1}{\left(j^{s-0.5}\sqrt{1+\epsilon_K}-j^{-0.5}\right)^2}\left(16K\ln 2 + \frac{2K(1+4\ln 2)}{j} + \frac{4}{j}\ln\frac{65536N}{K^3j(j+1)e^4} + \frac{2t}{j}\right) & \text{if } j > \left\lfloor \frac{\log_2 N}{K}\right\rfloor. \end{cases} \end{aligned}$$

The sequences $\{M_j\}_{j=1}^{\left\lfloor \frac{\log_2 N}{K}\right\rfloor -1}$ and $\{M_j\}_{j=\left\lfloor \frac{\log_2 N}{K}\right\rfloor +1}^{\left\lceil \frac{N}{K}\right\rceil}$ are decreasing sequences, since the numerators are decreasing sequences and the denominator is an increasing sequence whenever $s > 0.5$. When $K \leq \log_2 N$, we have

$$\begin{aligned} M \geq \max \Bigg(& \frac{1}{\left(\sqrt{1+\epsilon_K}-1\right)^2}\left(K(11+12\ln 2) + 4\ln\frac{N}{K(K+1)(2K+1)} + 2t\right), \\ & \frac{4K(1+\ln 8) + \frac{2K(1+4\ln 2)+4\ln\frac{256N}{K(\log_2 N+1)(\log_2 N+K+1)e^2}+2t}{\frac{\log_2 N}{K}}}{\left(\frac{\log_2 N}{K}^{s-0.5}\sqrt{1+\epsilon_K} - \frac{\log_2 N}{K}^{-0.5}\right)^2}, \\ & \frac{16K\ln 2 + \frac{2K(1+4\ln 2)+4\ln\frac{65536N}{K(\log_2 N+K)(\log_2 N+2K)e^4}+2t}{\frac{\log_2 N}{K}+1}}{\left(\left(\frac{\log_2 N}{K}+1\right)^{s-0.5}\sqrt{1+\epsilon_K} - \left(\frac{\log_2 N}{K}+1\right)^{-0.5}\right)^2}\Bigg). \end{aligned}$$

These three terms have sequentially smaller numerators and sequentially larger denominators, resulting in

$$M \geq \frac{1}{\left(\sqrt{1+\epsilon_K}-1\right)^2}\left(K(11+12\ln 2)+4\ln\frac{N}{K(K+1)(2K+1)}+2t\right).$$

When $K > \log_2 N$, the first two regimes of M_j are nonexistent, and so we have

$$M \geq \frac{1}{\left(\sqrt{1+\epsilon_K}-1\right)^2}\left(2K(1+12\ln 2)+4\ln\frac{32768N}{K^3e^4}+2t\right).$$

This completes the proof of Proposition 3. □

ACKNOWLEDGEMENTS

We thank Petros Boufounos and Mark Davenport for helpful discussions.

REFERENCES

[1] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego: Academic Press, 1999.

[2] D. L. Donoho, "Compressed sensing," *IEEE Trans. Info. Theory*, vol. 52, pp. 1289–1306, Sept. 2006.

[3] E. J. Candès, "Compressive sampling," in *Proc. International Congress of Mathematicians*, vol. 3, (Madrid, Spain), pp. 1433–1452, 2006.

[4] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 118–120, 124, July 2007.

[5] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of linear subspaces," *IEEE Trans. Info. Theory*, July 2007. Submitted.

[6] Y. M. Lu and M. N. Do, "Sampling signals from a union of subspaces," *IEEE Signal Processing Mag.*, vol. 25, pp. 41–47, Mar. 2008.

[7] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," Mar. 2008. Preprint.

[8] Y. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," 2008. Preprint.

[9] D. Baron, M. F. Duarte, S. Sarvotham, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing," 2005. Preprint.

[10] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, June 2008. To be published.

[11] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," July 2008. Preprint.

[12] M. S. Crouse, R. D. Nowak, and R. G. Baraniuk, "Wavelet-based statistical signal processing using Hidden Markov Models," *IEEE Trans. Signal Processing*, vol. 46, pp. 886–902, Apr. 1998.

[13] R. G. Baraniuk, "Fast reconstruction from incoherent projections." Workshop on Sparse Representations in Redundant Systems, May 2005.

[14] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, "Fast reconstruction of piecewise smooth signals from random projections," in *Proc. SPARS05*, (Rennes, France), Nov. 2005.

[15] C. La and M. N. Do, "Tree-based orthogonal matching pursuit algorithm for signal reconstruction," in *IEEE International Conference on Image Processing (ICIP)*, (Atlanta, GA), pp. 1277–1280, Oct. 2006.

[16] M. F. Duarte, M. B. Wakin, and R. G. Baraniuk, "Wavelet-domain compressive signal reconstruction using a hidden Markov tree model," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, (Las Vegas, NV), pp. 5137–5140, April 2008.

[17] K. Lee and Y. Bresler, "Selecting good Fourier measurements for compressed sensing." SIAM Conference on Imaging Science, July 2008.

[18] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and subgaussian ensembles," *Constructive Approximation*, Feb. 2008. To be published.

[19] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, vol. 20, p. 33, 1998.

[20] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Info. Theory*, vol. 52, pp. 4036–4048, Sept. 2006.

[21] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, Dec. 2007.

[22] J. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Info. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.

[23] D. L. Donoho, I. Drori, Y. Tsaig, and J. L. Starck, "Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit," 2006. Preprint.

[24] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing: Closing the gap between performance and complexity," Mar. 2008. Preprint.

[25] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Series I*, vol. 346, pp. 589–592, May 2008.

[26] R. G. Baraniuk and D. L. Jones, "A signal-dependent time-frequency representation: Fast algorithm for optimal kernel design," *IEEE Trans. Signal Processing*, vol. 42, pp. 134–146, Jan. 1994.

[27] R. Baraniuk, "Optimal tree approximation with wavelets," in *Wavelet Applications in Signal and Image Processing VII*, vol. 3813 of *Proc. SPIE*, (Denver, CO), pp. 196–207, July 1999.

[28] R. G. Baraniuk, R. A. DeVore, G. Kyriazis, and X. M. Yu, "Near best tree approximation," *Advances in Computational Mathematics*, vol. 16, pp. 357–373, May 2002.

[29] J. K. Romberg, H. Choi, and R. G. Baraniuk, "Bayesian tree-structured image modeling using wavelet-domain Hidden Markov Models," *IEEE Trans. Image Processing*, vol. 10, pp. 1056–1068, July 2001.

[30] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using a scale mixture of Gaussians in the wavelet domain," *IEEE Trans. Image Processing*, vol. 12, pp. 1338–1351, Nov. 2003.

[31] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[32] A. Cohen, W. Dahmen, I. Daubechies, and R. A. DeVore, "Tree approximation and optimal encoding," *Applied and Computational Harmonic Analysis*, vol. 11, pp. 192–226, Sept. 2001.

[33] D. Donoho, "CART and best ortho-basis: A connection," *Annals of Statistics*, vol. 25, pp. 1870–1911, Oct. 1997.

[34] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk, "Recovery of jointly sparse signals from few random projections," in *Proc. Workshop on Neural Info. Proc. Sys. (NIPS)*, (Vancouver), Nov. 2005.

[35] J. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing*, vol. 86, pp. 572–588, Apr. 2006.

[36] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, pp. 4203–4215, Dec. 2005.

[37] M. Ledoux, *The Concentration of Measure Phenomenon*. American Mathematical Society, 2001.

[38] G. G. Brown and B. O. Shubert, "On random binary trees," *Mathematics of Operations Research*, vol. 9, pp. 43–65, Feb. 1984.