Approximation Algorithms for Model-Based Compressive Sensing

Chinmay Hegde, Piotr Indyk, Ludwig Schmidt [*]

CSAIL, MIT

February 13, 2016

Abstract

Compressive Sensing (CS) states that a sparse signal can be recovered from a small number of linear measurements, and that this recovery can be performed efficiently in polynomial time. The framework of *model-based compressive sensing* (model-CS) leverages additional structure in the signal and provides new recovery schemes that can reduce the number of measurements even further. This idea has led to measurement-efficient recovery schemes for a variety of signal models. However, for any given model, model-CS requires an algorithm that solves the *model-projection problem*: given a query signal, report the signal in the model that is closest to the query signal. Often, this optimization problem can be computationally very expensive. Moreover, an *approximation* algorithm is not sufficient to provably recover the signal. As a result, the model-projection problem poses a fundamental obstacle for extending model-CS to many interesting classes of models.

In this paper, we introduce a new framework that we call *approximation-tolerant model-based compressive sensing*. This framework includes a range of algorithms for sparse recovery that require only *approximate solutions* for the model-projection problem. In essence, our work removes the aforementioned obstacle to model-based compressive sensing, thereby extending model-CS to a much wider class of signal models. Interestingly, all our algorithms involve both the minimization and maximization variants of the model-projection problem.

We instantiate our new framework for a new signal model that we call the Constrained Earth Mover Distance (CEMD) model. This model is particularly useful for signal ensembles where the positions of the nonzero coefficients do not change significantly as a function of spatial (or temporal) location. We develop novel approximation algorithms for both the maximization and the minimization versions of the model-projection problem via graph optimization techniques. Leveraging these algorithms and our framework results in a nearly sample-optimal sparse recovery scheme for the CEMD model.

1 Introduction

Over the last decade, a new *linear* approach for obtaining a succinct representation of n-dimensional vectors (or signals) has emerged. For any signal x, the representation is given by Ax, where A is an

[*] Authors listed in alphabetical order. The authors would like to thank Lei Hamilton, Chris Yu, Ligang Lu, and Detlef Hohl for helpful discussions. This work was supported in part by grants from the MITEI-Shell program, the MADALGO center, and the Packard Foundation. A conference version of this manuscript [1] appeared in the Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), held in January 2014.

$m \times n$ matrix, or possibly a random variable chosen from a suitable distribution over such matrices. The vector Ax is referred to as the *measurement vector* or *linear sketch* of x. Although m is usually chosen to be much smaller than n, the measurement vector Ax often contains plenty of useful information about the signal x.

A particularly useful and well-studied problem in this context is that of *robust sparse recovery.* A vector x is k-sparse if it has at most k non-zero coordinates. The robust sparse recovery problem is typically defined as follows: given the measurement vector $y = Ax + e$, where x is a k-sparse vector and e is the "noise" vector , find a signal estimate $\widehat{x}$ such that:

$$\|x - \widehat{x}\|_2 \; \leq \; C \cdot \|e\|_2 \, . \tag{1}$$

Sparse recovery has a tremendous number of applications in areas such as compressive sensing of signals [2, 3], genetic data analysis [4], and data stream algorithms [5, 6].

It is known that there exist matrices A and associated recovery algorithms that produce a signal estimate $\widehat{x}$ satisfying Equation (1) with a constant approximation factor C and number of measurements $m = O(k \log(n/k))$. It is also known that this bound on the number of measurements m is asymptotically *optimal* for some constant C; see [7] and [8] (building upon the classical results of [9, 10, 11]). The necessity of the "extra" logarithmic factor multiplying k is rather unfortunate: the quantity m determines the "compression rate", and for large n any logarithmic factor in m can worsen this rate tenfold.

On the other hand, more careful signal *modeling* offers a way to overcome the aforementioned limitation. Indeed, decades of research in signal processing have shown that not all signal supports (i.e., sets of non-zero coordinates) are equally common in practice. For example, in the case of certain time-domain signals such as signals transmitted by push-to-talk radios, the dominant coefficients of the signal tend to cluster together in contiguous "bursts". A formal approach to capture this additional *structure* is to assume that the support of the vector x belongs to a given family of supports $\mathbb{M}$, a so-called "model" (we say that x is $\mathbb{M}$-sparse). Note that the original k-sparse recovery problem corresponds to the particular case when the model $\mathbb{M}$ is the family of all k-subsets of $[n]$.

This modeling approach has several interesting ramifications, particularly in the context of robust sparse recovery. Recently, Baraniuk et al. provided a general framework called *model-based compressive sensing* [12]. For any "computationally tractable" and "small" family of supports, the scheme proposed in their work guarantees robust signal recovery with a nearly-optimal number of measurements $m = O(k)$, i.e., *without any logarithmic dependence on n.* Several other works have achieved similar performance gains both in theory and in practice; see, for example, [13, 14, 15, 16, 17].

While the model-based compressive sensing framework is general, it relies on two model-specific assumptions:

(1) *Model-based Restricted Isometry Property (RIP)*: The matrix A approximately preserves the ℓ_2-norm of all $\mathbb{M}$-sparse vectors.

(2) *Model projection oracle*: There exists an efficient algorithm that solves the *model-projection problem*: given an arbitrary vector x, the algorithm finds the $\mathbb{M}$-sparse vector x' that is closest to x, i.e., minimizes the ℓ_2-norm of the "tail" error $\|x - x'\|_2$.

By constructing matrices satisfying (1) and algorithms satisfying (2), researchers have developed robust signal recovery schemes for a wide variety of signal models, including block-sparsity [12], tree-sparsity [12], clustered sparsity [18], and separated spikes [19], to name a few.

Unfortunately, extending the model-based compressive sensing framework to more general models faces a significant obstacle. For the framework to apply, the model projection oracle has to be *exact* (i.e., the oracle finds the signal in the model with exactly minimal tail error). This fact may appear surprising, but in Section 3 we provide a *negative result* and prove that existing model-based recovery approaches fail to achieve the robust sparse recovery criterion (1) if the model projection oracle is not exact. Consequently, this burden of "exactness" excludes several useful design paradigms employed in *approximation algorithms*, i.e., algorithms which find a signal in the model that only approximately minimizes the tail error. A rich and extensive literature on approximation algorithms has emerged over the last 15 years, encompassing a variety of techniques such as greedy optimization, linear programming (LP) rounding, semidefinite programming (SDP) rounding, and Lagrangian relaxation. To the best of our knowledge, existing approaches for the model projection problem have largely focused on exact optimization techniques (e.g. dynamic programming [12, 20, 18], solving LPs without an integrality gap [19], etc.).

1.1 Summary of Our Results

In this paper, we introduce a new framework that we call *approximation-tolerant model-based compressive sensing.* This framework includes a range of algorithms for sparse recovery that require only *approximate solutions* for the model-projection problem. In essence, our work removes the aforementioned obstacle to model-based compressive sensing and therefore extends the framework to a much wider class of models. Simultaneously, our framework provides a principled approach to leverage the wealth of approximation algorithms for recovering structured sparse signals from linear measurements.

Instead of requiring one exact model projection oracle, our algorithms assume the existence of *two* oracles with complementary approximation guarantees: (i) Given $x \in \mathbb{R}^n$, a *tail approximation* oracle returns a support Ω_t in the model such that the norm of the tail $\|x - x_{\Omega_t}\|_2$ is approximately minimized. (ii) A *head approximation* oracle returns a support Ω_h in the model such that the norm of the head $\|x_{\Omega_h}\|_2$ is approximately maximized. Formally, we have:

$$\|x - x_{\Omega_t}\|_2 \;\leq\; c_T \cdot \min_{\Omega \in \mathbb{M}} \|x - x_\Omega\|_2 \qquad \text{and} \tag{2}$$

$$\|x_{\Omega_h}\|_2 \;\geq\; c_H \cdot \max_{\Omega \in \mathbb{M}} \|x_\Omega\|_2 \tag{3}$$

for some positive constants $c_H \leq 1$ and $c_T \geq 1$. Given access to these approximation oracles, we prove the following main result.

Theorem 1 (Signal recovery). *Consider a structured sparsity model $\mathcal{M} \subseteq \mathbb{R}^n$ and norm parameter $p \in \{1, 2\}$. Suppose that $x \in \mathcal{M}$ and that we observe m noisy linear measurements $y = Ax + e$. Suppose further that A satisfies the model-RIP in terms of the ℓ_p-norm, and that we are given access to head- and tail-approximation oracles $H(\cdot)$ and $T(\cdot)$ satisfying* (2) *and* (3)*, respectively. Then there exists an efficient algorithm that outputs a signal estimate $\widehat{x}$ such that $\|x - \widehat{x}\|_p \leq C\|e\|_p$ for some constant $C > 0$.*

We analyze the two cases $p = 1$ and $p = 2$ separately and develop two different types of recovery algorithms. The case of $p = 2$ is perhaps more well-studied in the literature and corresponds to the "standard" notion of the RIP. In this case, our recovery algorithms are extensions of IHT, CoSaMP, and their model-based counterparts [21, 22, 12]. The case of $p = 1$ has received attention in recent years and is applicable to the situation where the measurement matrix A is *itself* sparse. In this case, our recovery algorithms are extensions of those developed in [23, 24, 25]. For both types of algorithms, the sequence of signal estimates (x_k) produced by our algorithms exhibits *geometric convergence* to the true signal x, i.e., the norm of the error $\|x_k - x\|_p$ decreases by at least a constant factor in every iteration. The rate of convergence depends on the approximation constants c_T and c_H, as well as the RIP constants of the matrix A.

As a case study, we instantiate both the $p = 1$ and $p = 2$ cases in the context of the *Constrained Earth Mover's Distance* (CEMD) model introduced in [26]. In this model, the signal coefficients form an $h \times w$ grid and the support of each column has size at most s, for $n = h \cdot w$ and $k = s \cdot w$. For each pair of consecutive columns, say c and c', we define the Earth Mover's Distance (EMD) between them to be the minimum cost of matching the support sets of c and c' when viewed as point sets on a line. A signal support is said to belong to the CEMD model with "budget" B if the sum of all EMD distances between the consecutive columns is at most B. See Section 8 for a formal definition. Our framework leads to the first *nearly sample-optimal* recovery scheme for signals belonging to this model. The result is obtained by designing a novel head-approximation algorithm and proving approximation guarantees for the tail-approximation algorithm that was first described in [26].

1.2 Paper Outline

This paper includes the following contributions, organized by section. Before our contributions, we briefly review some background in Section 2.

A negative result for approximation oracles. In Section 3, we begin with the following negative result: combining an apprixmate model-projection oracle with the existing model-based compressive sensing approach of Baraniuk et al. [12] *does not suffice* to guarantee robust signal recovery for even the most trivial model. This serves as the motivation for a more sophisticated approach, which we develop throughout the rest of the paper.

Approximate model-iterative hard threholding (AM-IHT). In Section 4, we propose a new extension of the iterative hard thresholding (IHT) algorithm [22], which we call *approximate model iterative hard thresholding* (or AM-IHT). Informally, given head- and tail-approximation oracles and measurements $y = Ax + e$ with a matrix A satisfying the model-RIP, AM-IHT returns a signal estimate $\widehat{x}$ satisfying (1). We show that AM-IHT exhibits geometric convergence, and that the recovery guarantee for AM-IHT is asymptotically equivalent to the best available guarantees for model-based sparse recovery, despite using only approximate oracles.

Approximate model-CoSaMP (AM-CoSaMP). In Section 5, we propose a new extension of the compressive sampling matching pursuit algorithm (CoSaMP) [21], which we call *approximate model CoSaMP* (or AM-CoSAMP). As with AM-IHT, our proposed AM-CoSaMP algorithm requires a head-approximation oracle and a tail-approximation oracle. We show that AM-CoSaMP also exhibits geometric convergence, and that the recovery guarantee for AM-CoSaMP, as well as the RIP condition on A required for successful signal recovery, match the corresponding parameters for AM-IHT up to constant factors.

AM-IHT with sparse measurement matrices. In Section 6, we show that an approximation-tolerant approach similar to AM-IHT succeeds even when the measurement matrix A is *itself* sparse. Our approach leverages the notion of the restricted isometry property in the ℓ_1-norm, also called the *RIP-1*, which was first introduced in [23] and developed further in the model-based context by [27, 24, 25]. For sparse A, we propose a modification of AM-IHT, which we call *AM-IHT with RIP-1*. Our proposed algorithm also exhibits geometric convergence under the *model RIP-1* assumption on the measurement matrix A.

Compressive sensing with the CEMD Model. We design both head- and tail-approximation algorithms for the CEMD model: (i) Our tail-approximation oracle returns a support set with tail-approximation error at most a constant times larger than the optimal tail error. At the same time, the EMD-budget of the solution is still $O(B)$ (Theorem 34). (ii) Our head-approximation oracle returns a support set with head value at least a constant fraction of the optimal head value. Moreover, the EMD-budget of the solution is $O(B \log \frac{k}{w})$ (Theorem 26). Combining these algorithms into our new framework, we obtain a compressive sensing scheme for the CEMD model using $O(k \log(\frac{B}{k} \log(\frac{k}{w})))$ measurements for robust signal recovery. For a reasonable choice of parameters, e.g., $B = O(k)$, the bound specializes to $m = O(k \log\log(\frac{k}{w}))$, which is very close to the information-theoretic optimum of $m = O(k)$.

1.3 Prior Work

Prior to this paper, several efforts have been made to enable compressive sensing recovery for structured sparse signals with approximate projection oracles. The paper [28] discusses a Projected Landweber-type method that succeeds even when the projection oracle is approximate. However, the author assumes that the projection oracle provides an ϵ-*additive* tail approximation guarantee. In other words, for any given $x \in \mathbb{R}^n$, the model-approximation oracle returns a $\widehat{x} \in \mathcal{M}$ satisfying:

$$\|x - \widehat{x}\|_2 = \min_{x' \in \mathcal{M}} \left\|x - x'\right\|_2 + \varepsilon \tag{4}$$

for some parameter $\varepsilon > 0$. Under such conditions, there exists an algorithm that returns a signal within an $O(\epsilon)$-neighborhood of the optimal solution. However, approximation oracles that achieve low additive approximation guarantees satisfying (4) are rather rare.

On the other hand, the works [29, 30] assume the existence of a head-approximation oracle similar to our definition (3) and develop corresponding signal recovery algorithms. However, these approaches only provide signal recovery guarantees with an additive error term of $O(\|x_\Omega\|)$, where Ω is the set of the k largest coefficients in x. Therefore, this result is not directly comparable to our desired recovery guarantee (1).

Some more recent works have introduced the use of approximate projection oracles, albeit for a different type of signal model. There, the underlying assumption is that the signals of interest are sparse in a *redundant dictionary*. The paper [31] presents a sparse recovery algorithm for redundant dictionaries that succeeds with multiplicative approximation guarantees. However, their framework uses only the tail oracle and therefore is subject to the lower bound that we provide in Section 3. In particular, their guarantees make stringent assumptions on the maximum singular values of the sensing matrix A.

The paper [32] introduces an algorithm called *Signal Space CoSaMP* (SSCoSaMP), which also assumes the existence of multiplicative approximate oracles. However, the assumptions made on the oracles are restrictive. Interpreted in the model-based context, the oracles must capture a significant fraction of the optimal support in each iteration, which can be hard to achieve in practice. The more recent paper [33] proposes a version of SSCoSaMP which succeeds with oracles satisfying both multiplicative head- and tail-approximation guarantees. Indeed, our AM-CoSaMP algorithm and associated proofs are closely related to this work. However, AM-CoSaMP requires technically weaker conditions to succeed and our proof techniques are somewhat more concise. See Section 5 for a more detailed discussion on this topic.

In a parallel line of research, there have been several proposals for compressive sensing methods using *sparse* measurement matrices [23, 6]. Recent efforts have extended this line of work into the model-based setting. The paper [27] establishes both lower and upper bounds on the number of measurements required to satisfy the model RIP-1 for certain structured sparsity models. Assuming that the measurement matrix A satisfies the model RIP-1, the paper [25] proposes a modification of *expander iterative hard thresholding* (EIHT) [24], which achieves stable recovery for arbitrary structured sparsity models. As with the other algorithms for model-based compressive sensing, EIHT only works with *exact* model projection oracles. In Section 6, we propose a more general algorithm suitable for model-based recovery using only approximate projection oracles.

We instantiate our algorithmic results in the context of the Constrained Earth Mover's Distance (CEMD) model, developed in [26]. The model was originally motivated by the task of reconstructing time sequences of spatially sparse signals. There has been a substantial amount of work devoted to such signals, e.g., [34, 35]. We refer the reader to [26] for a more detailed discussion about the model and its applications. The paper introduced a tail oracle for the problem and empirically evaluated the performance of the recovery scheme. Although the use of the oracle was heuristic, the experiments demonstrate a substantial reduction in the number of measurements needed to recover slowly varying signals. In this paper, we provide a rigorous analysis of the tail-approximation oracle originally proposed in [26], as well as a novel head-approximation algorithm. Combining these two sub-routines yields a model-based compressive sensing scheme for the CEMD model using a nearly optimal number of measurements.

1.4 Subsequent Work

Since the appearance of the conference version of this manuscript [1], a number of works have explored some of its implications. The works [36, 37] develop approximation algorithms for the *tree-sparsity* model [12]. These algorithms, coupled with our framework, immediately imply sample-optimal recovery schemes for tree-sparse signals that run in *nearly linear-time.* Additionally, our approximation oracles for the CEMD model can be of independent interest in signal processing applications. For instance, [38] uses the tail-approximation procedure developed in Section 8.3 for detecting *faults* in subsurface seismic images. Investigations into further extensions are currently underway.

2 Preliminaries

We write $[n]$ to denote the set $\{1, 2, \ldots, n\}$ and $\mathcal{P}(A)$ to denote the power set of a set A. For a vector $x \in \mathbb{R}^n$ and a set $\Omega \subseteq [n]$, we write x_Ω for the restriction of x to Ω, i.e., $(x_\Omega)_i = x_i$ for $i \in \Omega$ and $(x_\Omega)_i = 0$ otherwise. Similarly, we write X_Ω for the submatrix of a matrix $X \in \mathbb{R}^{m \times n}$ containing the columns corresponding to Ω, i.e., a matrix in $\mathbb{R}^{m \times |\Omega|}$. Sometimes, we also restrict a matrix element-wise: for a set $\Omega \subseteq [m] \times [n]$, the matrix X_Ω is identical to X but the entries not contained in Ω are set to zero. The distinction between these two conventions will be clear from context.

A vector $x \in \mathbb{R}^n$ is said to be k-sparse if at most $k \leq n$ coordinates are nonzero. The support of x, $\operatorname{supp}(x) \subseteq [n]$, is the set of indices with nonzero entries in x. Hence $x_{\operatorname{supp}(x)} = x$. Observe that the set of *all* k-sparse signals is geometrically equivalent to the union of the $\binom{n}{k}$ canonical k-dimensional subspaces of $\mathbb{R}^n$. For a matrix $X \in \mathbb{R}^{h \times w}$, the support $\operatorname{supp}(X) \subseteq [h] \times [w]$ is also the set of indices corresponding to nonzero entries. For a matrix support set Ω, we denote the support of a column c in Ω with $\text{col-supp}(\Omega, c) = \{r \,|\, (r, c) \in \Omega\}$.

Often, some prior information about the support of a sparse signal x is available. A flexible way to model such prior information is to consider only the k-sparse signals with a permitted configuration of $\operatorname{supp}(x)$. This restriction motivates the notion of a *structured sparsity model*, which is geometrically equivalent to a subset of the $\binom{n}{k}$ canonical k-dimensional subspaces of $\mathbb{R}^n$.

Definition 2 (Structured sparsity model. From Definition 2 in [12])**.** *A structured sparsity model $\mathcal{M} \subseteq \mathbb{R}^n$ is the set of vectors $\mathcal{M} = \{x \in \mathbb{R}^n \,|\, \operatorname{supp}(x) \subseteq S$ for some $S \in \mathbb{M}\}$, where $\mathbb{M} = \{\Omega_1, \ldots, \Omega_l\}$ is the set of allowed structured supports with $\Omega_i \subseteq [n]$. We call $l = |\mathbb{M}|$ the size of the model $\mathcal{M}$.*

Note that the Ω_i in the definition above can have different cardinalities, but the largest cardinality will dictate the sample complexity in our bounds. Often it is convenient to work with the closure of $\mathbb{M}$ under taking subsets, which we denote with $\mathbb{M}^+ = \{\Omega \subseteq [n] \,|\, \Omega \subseteq S \text{ for some } S \in \mathbb{M}\}$. Then we can write the set of signals in the model as $\mathcal{M} = \{x \in \mathbb{R}^n \,|\, \operatorname{supp}(x) \in \mathbb{M}^+\}$.

In the analysis of our algorithms, we also use the notion of *model addition*: given two structured sparsity models $\mathcal{A}$ and $\mathcal{B}$, we define the sum $\mathcal{C} = \mathcal{A} \oplus \mathcal{B}$ as $\mathcal{C} = \{a + b \,|\, a \in \mathcal{A} \text{ and } b \in \mathcal{B}\}$ (i.e., the Minkowski sum). Similarly, we define the corresponding set of allowed supports as $\mathbb{C} = \mathbb{A} \oplus \mathbb{B} = \{\Omega \cup \Gamma \,|\, \Omega \in \mathbb{A} \text{ and } \Gamma \in \mathbb{B}\}$. We also use $\mathbb{C}^{\oplus t}$ as a shorthand for t-times addition, i.e., $\mathbb{C} \oplus \mathbb{C} \oplus \ldots \oplus \mathbb{C}$.

The framework of model-based compressive sensing [12] leverages the above notion of a structured sparsity model to design robust sparse recovery schemes. Specifically, the framework states that it is possible to recover a structured sparse signal $x \in \mathcal{M}$ from linear measurements $y = Ax + e$, provided that two conditions are satisfied: (i) the matrix A satisfies a variant of the restricted isometry property known as the *model-RIP*, and (ii) there exists an oracle that can efficiently *project* an arbitrary signal in $\mathbb{R}^n$ onto the model $\mathcal{M}$. We formalize these conditions as follows.

Definition 3 (Model-RIP. From Definition 3 in [12])**.** *The matrix $A \in \mathbb{R}^{m \times n}$ has the $(\delta, \mathbb{M})$-model-RIP if the following inequalities hold for all x with $\operatorname{supp}(x) \in \mathbb{M}^+$:*

$$(1 - \delta)\|x\|_2^2 \ \leq \ \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2 \,. \tag{5}$$

The following properties are direct consequences of the model-RIP and will prove useful in our proofs in Sections 4 and 5.

Fact 4 (adapted from Section 3 in [21])**.** *Let $A \in \mathbb{R}^{m \times n}$ be a matrix satisfying the $(\delta, \mathbb{M})$-model-RIP. Moreover, let Ω be a support in the model, i.e., $\Omega \in \mathbb{M}^+$. Then the following properties hold for all $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$:*

$$\begin{aligned} \left\|A_\Omega^T y\right\|_2 &\leq \sqrt{1 + \delta}\,\|y\|_2 \,, \\ \left\|A_\Omega^T A_\Omega x\right\|_2 &\leq (1 + \delta)\|x\|_2 \,, \\ \left\|(I - A_\Omega^T A_\Omega) x\right\|_2 &\leq \delta \|x\|_2 \,. \end{aligned}$$

Definition 5 (Model-projection oracle. From Section 3.2 in [12]). *A model-projection oracle is a function $M : \mathbb{R}^n \to \mathcal{P}([n])$ such that the following two properties hold for all $x \in \mathbb{R}^n$.*
Output model sparsity: $M(x) \in \mathbb{M}^+$.
Optimal model projection: Let $\Omega' = M(x)$. Then $\|x - x_{\Omega'}\|_2 = \min_{\Omega \in \mathbb{M}} \|x - x_\Omega\|_2$.

Sometimes, we use a model-projection oracle M as a function from $\mathbb{R}^n$ to $\mathbb{R}^n$. This can be seen as a simple extension of Definition 5 where $M(x) = x_\Omega$, $\Omega = M'(x)$, and M' satisfies Definition 5.

Under these conditions, the authors of [12] show that compressive sampling matching pursuit (CoSaMP [21]) and iterative hard thresholding (IHT [22]) — two popular algorithms for sparse recovery — can be modified to achieve robust sparse recovery for the model $\mathcal{M}$. In particular, the modified version of IHT (called *Model-IHT* [12]) executes the following iterations until convergence:

$$x^{i+1} \leftarrow M(x^i + A^T(y - Ax^i)) \,, \tag{6}$$

where $x^1 = 0$ is the initial signal estimate. From a sampling complexity perspective, the benefit of this approach stems from the model-RIP assumption. Indeed, the following result indicates that with high probability, a large class of measurement matrices A satisfies the model-RIP with a *nearly optimal* number of rows:

Fact 6 ([39, 12]). *Let $\mathbb{M}$ be a structured sparsity model and let k be the size of the largest support in the model, i.e., $k = \max_{\Omega \in \mathbb{M}} |\Omega|$. Let $A \in \mathbb{R}^{m \times n}$ be a matrix with i.i.d. sub-Gaussian entries. Then there is a constant c such that for $0 < \delta < 1$, any $t > 0$, and*

$$m \;\geq\; \frac{c}{\delta^2} \left(k \log \frac{1}{\delta} + \log|\mathbb{M}| + t \right) \,,$$

A has the $(\delta, \mathbb{M})$-model-RIP with probability at least $1 - e^{-t}$.

Since δ and t are typically constants, this bound can often be summarized as

$$m \;=\; O(k + \log|\mathbb{M}|) \,.$$

If the number of permissible supports (or equivalently, subspaces) $|\mathbb{M}|$ is asymptotically smaller than $\binom{n}{k}$, then m can be smaller than the $O(k \log \frac{n}{k})$ measurement bound from "standard" compressive sensing. In the ideal case, we have $m = \mathrm{poly}(n) \cdot 2^{O(k)}$, which implies a measurement bound of $m = O(k)$ under the very mild assumption that $k = \Omega(\log n)$. Since $m = k$ measurements are necessary to reconstruct any k-sparse signal, this asymptotic behavior of m is information-theoretically optimal up to constant factors.

While model-based recovery approaches improve upon "standard" sparsity-based approaches in terms of sample-complexity, the computational cost of signal recovery crucially depends on the model-projection oracle M. Observe that Model-IHT (Equation 6) involves one invocation of the model-projection oracle M per iteration, and hence its overall running time scales with that of M. Therefore, model-based recovery approaches are relevant *only* in situations where efficient algorithms for finding the optimal model-projection are available.

3 A Negative Result

For many structured sparsity models, computing an optimal model-projection can be a challenging task. One way to mitigate this computational burden is to use *approximate* model-projection oracles, i.e., oracles that solve the model-projection problem only approximately. However, in this section we show that such oracles cannot be integrated into Model-IHT (Equation 6) in a straightforward manner.

Consider the standard compressive sensing setting, where the "model" consists of the set of all k-sparse signals. Of course, finding the optimal model projection in this case is simple: for any signal x, the oracle $T_k(\cdot)$ returns the k largest coefficients of x in terms of absolute value. But for illustrative purposes, let us consider a slightly different oracle that is approximate in the following sense. Let c be an arbitrary constant and let T'_k be a projection oracle such that for any $a \in \mathbb{R}^n$ we have:

$$\left\|a - T'_k(a)\right\|_2 \leq c\|a - T_k(a)\|_2 \,. \tag{7}$$

We show that we can construct an "adversarial" approximation oracle T'_k that always returns $T'_k(a) = 0$ but still satisfies (7) for all signals a encountered during the execution of Model-IHT. In particular, we use this oracle in Model-IHT and start with the initial signal estimate $x^0 = 0$. We will show that such an adversarial oracle still satisfies (7) for the first iteration of Model-IHT. As a result, Model-IHT with this adversarial oracle remains stuck at the zero signal estimate and cannot recover the true signal.

Recall that Model-IHT with projection oracle T'_k iterates

$$x^{i+1} \leftarrow T'_k(x^i + A^T(y - Ax^i)) \,, \tag{8}$$

which in the first iteration gives

$$x^1 \leftarrow T'_k(A^T y) \,.$$

Consider the simplest case where the signal x is 1-sparse with $x_1 = 1$ and $x_i = 0$ for $i \neq 1$, i.e., $x = e_1$. Given a measurement matrix A with $(\delta, O(1))$-RIP for small δ, Model-IHT needs to perfectly recover x from Ax. It is known that random matrices $A \in \mathbb{R}^{m \times n}$ with $A_{i,j} = \pm 1/\sqrt{m}$ chosen i.i.d. uniformly at random satisfy this RIP for $m = O(\log n)$ with high probability [40].[1] We prove that our "adversarial" oracle $T'_k(a) = 0$ satisfies the approximation guarantee (7) for its input $a = A^T y = A^T A e_1$ with high probability. Hence, $x^1 = x^0 = 0$ and Model-IHT cannot make progress. Intuitively, the tail $a - T_k(a)$ contains so much "noise" that the adversarial approximation oracle T'_k does not need to find a good sparse support for a and can simply return a signal estimate of 0.

Consider the components of the vector $a = A^T A e_1$: a_i is the inner product of the first column of A with the i-th column of A. Clearly, we have $a_1 = 1$ and $-1 \leq a_i \leq 1$ for $i \neq 1$. Therefore, $T_k(a) = e_1$ is an optimal projection and $\|a - T_k(a)\|_2^2 = \|a\|_2^2 - 1$. In order to show that the adversarial oracle $T'_k(a)$ satisfies the guarantee (7) with constant c, we need to prove that:

$$\|a\|_2^2 \leq c^2(\|a\|_2^2 - 1) \,.$$

Therefore, it suffices to show that $\|a\|_2^2 \geq \frac{c^2}{c^2-1}$. Observe that $\|a\|_2^2 = 1 + \sum_{i=2}^n a_i^2$, where the a_i are independent. For $i \neq 1$, each a_i is the sum of m independent $\pm\frac{1}{m}$ random variables (with $p = 1/2$) and so $\mathbb{E}[a_i^2] = \frac{1}{m}$. We can use Hoeffding's inequality to show that $\sum_{i=2}^n a_i^2$ does not deviate from its mean $\frac{n-1}{m}$ by more than $O(\sqrt{n \log n})$ with high probability. Since $m = O(\log n)$, this shows that for any constant $c > 1$, we will have

$$\|a\|_2^2 = 1 + \sum_{i=2}^n a_i^2 \geq \frac{c^2}{c^2 - 1}$$

[1] These are the so-called *Rademacher* matrices.

with high probability for sufficiently large n.

Therefore, we have shown that (8) does *not* result in a model-based signal recovery algorithm with provable convergence to the correct result x. In the rest of this paper, we develop several alternative approaches that do achieve convergence to the correct result while using approximate projection-oracles.

4 Approximate Model-IHT

We now introduce our *approximation-tolerant model-based compressive sensing* framework. Essentially, we extend the model-based compressive sensing framework to work with approximate projection oracles, which we formalize in the definitions below. This extension enables model-based compressive sensing in cases where optimal model projections are beyond our reach, but approximate projections are still efficiently computable.

The core idea of our framework is to utilize *two* different notions of approximate projection oracles, defined as follows.

Definition 7 (Head approximation oracle). *Let $\mathbb{M}, \mathbb{M}_H \subseteq \mathcal{P}([n])$, $p \geq 1$, and $c_H \in \mathbb{R}$. Then $H : \mathbb{R}^n \to \mathcal{P}([n])$ is a $(c_H, \mathbb{M}, \mathbb{M}_H, p)$-head-approximation oracle if the following two properties hold for all $x \in \mathbb{R}^n$:*
Output model sparsity: $H(x) \in \mathbb{M}_H^+$.
Head approximation: Let $\Omega' = H(x)$. Then $\|x_{\Omega'}\|_p \geq c_H \|x_\Omega\|_p$ for all $\Omega \in \mathbb{M}$.

Definition 8 (Tail approximation oracle). *Let $\mathbb{M}, \mathbb{M}_T \subseteq \mathcal{P}([n])$, $p \geq 1$ and $c_T \in \mathbb{R}$. Then $T : \mathbb{R}^n \to \mathcal{P}([n])$ is a $(c_T, \mathbb{M}, \mathbb{M}_T, p)$-tail-approximation oracle if the following two properties hold for all $x \in \mathbb{R}^n$:*
Output model sparsity: $T(x) \in \mathbb{M}_T^+$.
Tail approximation: Let $\Omega' = T(x)$. Then $\|x - x_{\Omega'}\|_p \leq c_T \|x - x_\Omega\|_p$ for all $\Omega \in \mathbb{M}$.

We trivially observe that a head approximation oracle with approximation factor $c_H = 1$ is equivalent to a tail approximation oracle with factor $c_T = 1$, and vice versa. Further, we observe that for any model $\mathcal{M}$, if $x \in \mathcal{M}$ then $\|x - x_\Omega\|_2 = 0$ for some $\Omega \in \mathbb{M}$. Hence, *any* tail approximation oracle must be exact in the sense that the returned support Ω' has to satisfy $\|x - x_{\Omega'}\|_2 = 0$, or equivalently, $\mathrm{supp}(x) \subseteq T(x)$. On the other hand, we note that $H(x)$ does not need to return an optimal support if the input signal x is in the model $\mathcal{M}$.

An important feature of the above definitions of approximation oracles is that they permit projections into *larger* models. In other words, the oracle can potentially return a signal that belongs to a larger model $\mathbb{M}' \supseteq \mathbb{M}$. For example, a tail-approximation oracle for the CEMD model with parameters (k, B) is allowed to return a signal with parameters $(2k, 2B)$, thereby relaxing both the sparsity constraint and the EMD-budget. We exploit this feature in our algorithms in Section 8.

Equipped with these notions of approximate projection oracles, we introduce a new algorithm for model-based compressive sensing. We call our algorithm *Approximate Model-IHT* (AM-IHT); see Algorithm 1 for a full description. Notice that every iteration of AM-IHT uses *both* a head-approximation oracle H and a tail-approximation oracle T. This is in contrast to the Model-IHT algorithm discussed above in Section 3, which solely made use of a tail approximation oracle T'.

Our main result of this section (Theorem 11) states the following: if the measurement matrix A satisfies the model-RIP for $\mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H$ and approximate projection oracles H and T are available,

Algorithm 1 Approximate Model-IHT

```
function AM-IHT($y, A, t$)
    $x^0 \leftarrow 0$
    for $i \leftarrow 0, \ldots, t$ do
        $b^i \leftarrow A^T(y - Ax^i)$
        $x^{i+1} \leftarrow T(x^i + H(b^i))$
    return $x^{t+1}$
```

then AM-IHT exhibits provably robust recovery. We make the following assumptions in the analysis of AM-IHT: (i) $x \in \mathbb{R}^n$ and $x \in \mathcal{M}$. (ii) $y = Ax + e$ for an arbitrary $e \in \mathbb{R}^m$ (the measurement noise). (iii) T is a $(c_T, \mathbb{M}, \mathbb{M}_T, 2)$-tail-approximation oracle. (iv) H is a $(c_H, \mathbb{M}_T \oplus \mathbb{M}, \mathbb{M}_H, 2)$-head-approximation-oracle. (v) A has the $(\delta, \mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H)$-model-RIP.

As in IHT, we use the *residual proxy* $b^i = A^T(y - Ax^i)$ as the update in each iteration (see Algorithm 1). The key idea of our proof is the following: when applied to the residual proxy b^i, the head-approximation oracle H returns a support Γ that contains "most" of the relevant mass contained in r^i. Before we formalize this statement in Lemma 10, we first establish the RIP of A on all relevant vectors.

Lemma 9. *Let $r^i = x - x^i$, $\Omega = \operatorname{supp}(r^i)$, and $\Gamma = \operatorname{supp}(H(b^i))$. For all $x' \in \mathbb{R}^n$ with $\operatorname{supp}(x') \subseteq \Omega \cup \Gamma$ we have*

$$(1-\delta)\|x'\|_2^2 \le \|Ax'\|_2^2 \le (1+\delta)\|x'\|_2^2 \,.$$

Proof. By the definition of T, we have $\operatorname{supp}(x^i) \in \mathbb{M}_T$. Since $\operatorname{supp}(x) \in \mathbb{M}$, we have $\operatorname{supp}(x - x^i) \in \mathbb{M}_T \oplus \mathbb{M}$ and hence $\Omega \in \mathbb{M}_T \oplus \mathbb{M}$. Moreover, $\operatorname{supp}(H(b^i)) \in \mathbb{M}_H$ by the definition of H. Therefore $\Omega \cup \Gamma \in \mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H$, which allows us to use the model-RIP of A on x' with $\operatorname{supp}(x') \subseteq \Omega \cup \Gamma$. □

We now establish our main lemma, which will also prove useful in Section 5. A similar result (with a different derivation approach and different constants) appears in Section 4 of the conference version of this manuscript [1].

Lemma 10. *Let $r^i = x - x^i$ and $\Gamma = \operatorname{supp}(H(b^i))$. Then,*

$$\|r^i_{\Gamma^c}\|_2 \le \sqrt{1-\alpha_0^2}\,\|r^i\|_2 + \left[\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}}\right]\|e\|_2 \,. \tag{9}$$

where

$$\alpha_0 = c_H(1-\delta) - \delta \qquad \textit{and} \qquad \beta_0 = (1+c_H)\sqrt{1+\delta}\,.$$

We assume that c_H and δ are such that $\alpha_0 > 0$.

Proof. We provide lower and upper bounds on $\|H(b^i)\|_2 = \|b^i_\Gamma\|_2$, where $b^i = A^T(y - Ax_i) = A^TAr^i + A^Te$. Let $\Omega = \operatorname{supp}(r^i)$. From the head-approximation property, we can bound $\|b^i_\Gamma\|_2$ as:

$$\begin{aligned}
\|b^i_\Gamma\|_2 &= \|A_\Gamma^T A r^i + A_\Gamma^T e\|_2 \\
&\ge c_H \|A_\Omega^T A r^i + A_\Omega^T e\|_2 \\
&\ge c_H \|A_\Omega^T A_\Omega r^i\|_2 - c_H \|A_\Omega^T e\|_2 \\
&\ge c_H(1-\delta)\|r^i\|_2 - c_H\sqrt{1+\delta}\,\|e\|_2 \,,
\end{aligned}$$

where the inequalities follow from Fact 4 and the triangle inequality. This provides the lower bound on $\left\|b^i_\Gamma\right\|_2$.

Now, consider r_Γ. By repeated use of the triangle inequality, we get

$$\begin{aligned}
\left\|b^i_\Gamma\right\|_2 &= \left\|A^T_\Gamma A r^i + A^T_\Gamma e\right\|_2 \\
&= \left\|A^T_\Gamma A r^i - r^i_\Gamma + r^i_\Gamma + A^T_\Gamma e\right\|_2 \\
&\leq \left\|A^T_\Gamma A r^i - r^i_\Gamma\right\|_2 + \left\|r^i_\Gamma\right\|_2 + \left\|A^T_\Gamma e\right\|_2 \\
&\leq \left\|A^T_{\Gamma\cup\Omega} A r^i - r^i_{\Gamma\cup\Omega}\right\|_2 + \left\|r^i_\Gamma\right\|_2 + \sqrt{1+\delta}\|e\|_2 \\
&\leq \delta\left\|r^i\right\|_2 + \left\|r^i_\Gamma\right\|_2 + \sqrt{1+\delta}\|e\|_2 \,,
\end{aligned}$$

where the last inequality again follows from Fact 4. This provides the upper bound on $\left\|b^i_\Gamma\right\|_2$.

Combining the two bounds and grouping terms, we obtain the following inequality. In order to simplify notation, we write $\alpha_0 = c_H(1-\delta) - \delta$ and $\beta_0 = (1+c_H)\sqrt{1+\delta}$.

$$\left\|r^i_\Gamma\right\|_2 \geq \alpha_0\left\|r^i\right\|_2 - \beta_0\|e\|_2 \,. \tag{10}$$

Next, we examine the right hand side of (10) more carefully. Let us assume that the RIP constant δ is set to be small enough such that it satisfies $c_H > \delta/(1-\delta)$. There are two mutually exclusive cases:

Case 1: The value of $\|r^i\|_2$ satisfies $\alpha_0\|r^i\|_2 \leq \beta_0\|e\|_2$. Then, consider the vector $r^i_{\Gamma^c}$, i.e., the vector r^i restricted to the set of coordinates in the complement of Γ. Clearly, its norm is smaller than $\left\|r^i\right\|_2$. Therefore, we have

$$\left\|r^i_{\Gamma^c}\right\|_2 \leq \frac{\beta_0}{\alpha_0}\|e\|_2 \,. \tag{11}$$

Case 2: The value of $\|r^i\|_2$ satisfies $\alpha_0\|r^i\|_2 \geq \beta_0\|e\|_2$. Rewriting (10), we get

$$\left\|r^i_\Gamma\right\|_2 \geq \left\|r^i\right\|_2\left(\alpha_0 - \frac{\beta_0\|e\|_2}{\|r_i\|_2}\right) \,.$$

Moreover, we also have $\left\|r^i\right\|^2_2 = \left\|r^i_\Gamma\right\|^2_2 + \left\|r^i_{\Gamma^c}\right\|^2_2$. Therefore, we obtain

$$\left\|r^i_{\Gamma^c}\right\|_2 \leq \left\|r^i\right\|_2\sqrt{1 - \left(\alpha_0 - \beta_0\frac{\|e\|_2}{\left\|r^i\right\|_2}\right)^2} \,. \tag{12}$$

We can simplify the right hand side using the following geometric argument, adapted from [41]. Denote $\omega_0 = \alpha_0 - \beta_0\|e\|_2/\left\|r^i\right\|_2$. Then, $0 \leq \omega_0 < 1$ because $\alpha_0\|r^i\|_2 \geq \beta_0\|e\|_2$, $\alpha_0 < 1$, and $\beta_0 \geq 1$. The function $g(\omega_0) = \sqrt{1-\omega_0^2}$ traces an arc of the unit circle as a function of ω_0 and therefore is upper-bounded by the y-coordinate of *any* tangent line to the circle evaluated at ω_0. For a free parameter $0 < \omega < 1$ (the tangent point of the tangent line), a straightforward calculation yields that

$$\sqrt{1-\omega_0^2} \leq \frac{1}{\sqrt{1-\omega^2}} - \frac{\omega}{\sqrt{1-\omega^2}}\omega_0 \,.$$

Therefore, substituting into the bound for $\left\|r^i_{\Gamma^c}\right\|_2$, we get:

$$\begin{aligned}
\left\|r^i_{\Gamma^c}\right\|_2 &\leq \left\|r^i\right\|_2\left(\frac{1}{\sqrt{1-\omega^2}} - \frac{\omega}{\sqrt{1-\omega^2}}\left(\alpha_0 - \beta_0\frac{\|e\|_2}{\left\|r^i\right\|_2}\right)\right) \\
&= \frac{1-\omega\alpha_0}{\sqrt{1-\omega^2}}\left\|r^i\right\|_2 + \frac{\omega\beta_0}{\sqrt{1-\omega^2}}\|e\|_2 \,.
\end{aligned}$$

The coefficient preceding $\left\|r^i\right\|_2$ determines the overall convergence rate, and the minimum value of the coefficient is attained by setting $\omega = \alpha_0$. Substituting, we obtain

$$\left\|r^i_{\Gamma^c}\right\|_2 \leq \sqrt{1-\alpha_0^2}\left\|r^i\right\|_2 + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}}\|e\|_2 . \tag{13}$$

Combining the mutually exclusive cases (11) and (13), we obtain

$$\left\|r^i_{\Gamma^c}\right\|_2 \leq \sqrt{1-\alpha_0^2}\left\|r^i\right\|_2 + \left[\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}}\right]\|e\|_2 ,$$

which proves the lemma. □

Theorem 11 (Geometric convergence of AM-IHT)**.** *Let $r^i = x - x^i$, where x^i is the signal estimate computed by AM-IHT in iteration i. Then,*

$$\left\|r^{i+1}\right\|_2 \leq \alpha\left\|r^i\right\|_2 + \beta\|e\|_2 ,$$

where

$$\alpha = (1+c_T)\left[\delta + \sqrt{1-\alpha_0^2}\right], \qquad \beta = (1+c_T)\left[\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}} + \sqrt{1+\delta}\right],$$
$$\alpha_0 = c_H(1-\delta) - \delta , \qquad \beta_0 = (1+c_H)\sqrt{1+\delta} .$$

We assume that c_H and δ are such that $\alpha_0 > 0$.

Proof. Let $a = x^i + H(b^i)$. From the triangle inequality, we have:

$$\begin{aligned}
\left\|x - x^{i+1}\right\|_2 &= \|x - T(a)\|_2 \\
&\leq \|x-a\|_2 + \|a - T(a)\|_2 \\
&\leq (1+c_T)\|x-a\|_2 \\
&= (1+c_T)\left\|x - x^i - H(b^i)\right\|_2 \\
&= (1+c_T)\left\|r^i - H(A^TAr^i + A^Te)\right\|_2 .
\end{aligned} \tag{14}$$

We can further bound $\left\|r^i - H(A^TAr^i + A^Te)\right\|_2$ in terms of $\left\|r^i\right\|_2$. Let $\Omega = \operatorname{supp}(r^i)$ and $\Gamma = \operatorname{supp}(H(A^TAr^i + A^Te))$. We have the inequalities

$$\begin{aligned}
\left\|r^i - H(A^TAr^i + A^Te)\right\|_2 &= \left\|r^i_\Gamma + r^i_{\Gamma^c} - A^T_\Gamma Ar^i + A^T_\Gamma e\right\|_2 \\
&\leq \left\|A^T_\Gamma Ar^i - r^i_\Gamma\right\|_2 + \left\|r^i_{\Gamma^c}\right\|_2 + \left\|A^T_\Gamma e\right\|_2 \\
&\leq \left\|A^T_{\Gamma\cup\Omega} Ar^i - r^i_{\Gamma\cup\Omega}\right\|_2 + \left\|r^i_{\Gamma^c}\right\|_2 + \left\|A^T_\Gamma e\right\|_2 \\
&\leq \delta\left\|r^i\right\|_2 + \sqrt{1-\alpha_0^2}\left\|r^i\right\|_2 + \left[\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}} + \sqrt{1+\delta}\right]\|e\|_2 ,
\end{aligned}$$

where the last inequality follows from the RIP and (9). Putting this together with (14) and grouping terms, we get

$$\left\|x - x^{i+1}\right\|_2 \leq \alpha\left\|x - x^i\right\|_2 + \beta\|e\|_2 , \tag{15}$$

thus proving the Theorem. □

In the noiseless case, we can ignore the second term and only focus on the leading recurrence factor:

$$\alpha = (1 + c_T) \left(\delta + \sqrt{1 - (c_H(1 - \delta) - \delta)^2} \right) .$$

For convergence, we need α to be strictly smaller than 1. Note that we can make δ as small as we desire since this assumption only affects the measurement bound by a constant factor. Therefore, the following condition must hold for guaranteed convergence:

$$(1 + c_T)\sqrt{1 - c_H^2} < 1 \, , \qquad \text{or equivalently,} \qquad c_H^2 > 1 - \frac{1}{(1 + c_T)^2} \, . \tag{16}$$

Under this condition, AM-IHT exhibits geometric convergence comparable to the existing model-based compressive sensing results of [12]. AM-IHT achieves this *despite using only approximate projection oracles.* In Section 7, we relax condition (16) so that geometric convergence is possible for *any* constants c_T and c_H.

The geometric convergence of AM-IHT implies that the algorithm quickly recovers a good signal estimate. Formally, we obtain:

Corollary 12. *Let T and H be approximate projection oracles with c_T and c_H such that $0 < \alpha < 1$. Then after $t = \left\lceil \frac{\log \frac{\|x\|_2}{\|e\|_2}}{\log \frac{1}{\alpha}} \right\rceil$ iterations, AM-IHT returns a signal estimate $\widehat{x}$ satisfying*

$$\|x - \widehat{x}\|_2 \leq \left(1 + \frac{\beta}{1 - \alpha}\right) \|e\|_2 \, .$$

Proof. As before, let $r^i = x - x^i$. Using $\|r^0\|_2 = \|x\|_2$, Theorem 11, and a simple inductive argument shows that

$$\|r^{i+1}\|_2 \leq \alpha^i \|x\|_2 + \beta \|e\|_2 \sum_{j=0}^{i} \alpha^j \, .$$

For $i = \left\lceil \frac{\log \frac{\|x\|_2}{\|e\|_2}}{\log \frac{1}{\alpha}} \right\rceil$, we get $\alpha^i \|x\|_2 \leq \|e\|_2$. Moreover, we can bound the geometric series $\sum_{j=0}^{t} \alpha^j$ by $\frac{1}{1-\alpha}$. Combining these bounds gives the guarantee stated in the theorem. □

5 Approximate Model-CoSaMP

In this Section, we propose a second algorithm for model-based compressive sensing with approximate projection oracles. Our algorithm is a generalization of model-based CoSaMP, which was initially developed in [12]. We call our variant *Approximate Model-CoSaMP* (or AM-CoSaMP); see Algorithm 2 for a complete description.

Algorithm 2 closely resembles the *Signal-Space CoSaMP* (or SSCoSaMP) algorithm proposed and analyzed in [32, 33]. Like our approach, SSCoSaMP also makes assumptions about the existence of head- and tail-approximation oracles. However, there are some important technical differences in our development. SSCoSaMP was introduced in the context of recovering signals that are sparse in overcomplete and incoherent dictionaries. In contrast, we focus on recovering signals from structured sparsity models.

Algorithm 2 Approximate Model-CoSaMP

```
function AM-CoSaMP($y, A, t$)
    $x^0 \leftarrow 0$
    for $i \leftarrow 0, \ldots, t$ do
        $b^i \leftarrow A^T(y - Ax^i)$
        $\Gamma \leftarrow \mathrm{supp}(H(b^i))$
        $S \leftarrow \Gamma \cup \mathrm{supp}(x^i)$
        $z|_S \leftarrow A_S^{\dagger} y, \quad z|_{S^C} \leftarrow 0$
        $x^{i+1} \leftarrow T(z)$
    return $x^{t+1}$
```

Moreover, the authors of [32, 33] assume that a *single* oracle simultaneously achieves the conditions specified in Definitions 7 and 8. In contrast, our approach assumes the existence of two separate head- and tail-approximation oracles and consequently is somewhat more general. Finally, our analysis is simpler and more concise than that provided in [32, 33] and follows directly from the results in Section 4.

We prove that AM-CoSaMP (Alg. 2) exhibits robust signal recovery. We make the same assumptions as in Section 4: (i) $x \in \mathbb{R}^n$ and $x \in \mathcal{M}$. (ii) $y = Ax + e$ for an arbitrary $e \in \mathbb{R}^m$ (the measurement noise). (iii) T is a $(c_T, \mathbb{M}, \mathbb{M}_T, 2)$-tail-approximation oracle. (iv) H is a $(c_H, \mathbb{M}_T \oplus \mathbb{M}, \mathbb{M}_H, 2)$-head-approximation-oracle. (v) A has the $(\delta, \mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H)$-model-RIP. Our main result in this section is the following:

Theorem 13 (Geometric convergence of AM-CoSaMP)**.** *Let $r^i = x - x^i$, where x^i is the signal estimate computed by AM-CoSaMP in iteration i. Then,*

$$\|r^{i+1}\|_2 \leq \alpha \|r^i\|_2 + \beta \|e\|_2 \,,$$

where

$$\begin{aligned}
\alpha &= (1 + c_T)\sqrt{\frac{1+\delta}{1-\delta}}\sqrt{1-\alpha_0^2}\,, \\
\beta &= (1 + c_T)\left[\sqrt{\frac{1+\delta}{1-\delta}}\left(\frac{\beta_0}{\alpha_0} + \frac{\alpha_0\beta_0}{\sqrt{1-\alpha_0^2}}\right) + \frac{2}{\sqrt{1-\delta}}\right], \\
\alpha_0 &= c_H(1-\delta) - \delta\,, \\
\beta_0 &= (1 + c_H)\sqrt{1+\delta}\,.
\end{aligned}$$

Proof. We can bound the error $\|r^{i+1}\|_2$ as follows:

$$\begin{aligned}\left\|r^{i+1}\right\|_2 &= \left\|x - x^{i+1}\right\|_2 \\ &\leq \left\|x^{i+1} - z\right\|_2 + \|x - z\|_2 \\ &\leq c_T\|x - z\|_2 + \|x - z\|_2 \\ &= (1 + c_T)\|x - z\|_2 \\ &\leq (1 + c_T)\frac{\|A(x - z)\|_2}{\sqrt{1-\delta}} \\ &= (1 + c_T)\frac{\|Ax - Az\|_2}{\sqrt{1-\delta}} .\end{aligned}$$

Most of these inequalities follow the same steps as the proof provided in [21]. The second relation above follows from the triangle inequality, the third relation follows from the tail approximation property and the fifth relation follows from the $(\delta, \mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H)$-model-RIP of A.

We also have $Ax = y - e$ and $Az = A_S z_S$. Substituting, we get:

$$\begin{aligned}\left\|r^{i+1}\right\|_2 &\leq (1 + c_T)\left(\frac{\|y - A_S z_S\|_2}{\sqrt{1-\delta}} + \frac{\|e\|_2}{\sqrt{1-\delta}}\right) \\ &\leq (1 + c_T)\left(\frac{\|y - A_S x_S\|_2}{\sqrt{1-\delta}} + \frac{\|e\|_2}{\sqrt{1-\delta}}\right) . \qquad (17)\end{aligned}$$

The first inequality follows from the triangle inequality and the second from the fact that z_S is the least squares estimate $A_S^{\dagger} y$ (in particular, it is at least as good as x_S).

Now, observe that $y = Ax + e = A_S x_S + A_{S^c} x_{S^c} + e$. Therefore, we can further simplify inequality (17) as

$$\begin{aligned}\left\|r^{i+1}\right\|_2 &\leq (1 + c_T)\frac{\|A_{S^c} x_{S^c}\|_2}{\sqrt{1-\delta}} + (1 + c_T)\frac{2\|e\|_2}{\sqrt{1-\delta}} \\ &\leq (1 + c_T)\frac{\sqrt{1+\delta}}{\sqrt{1-\delta}}\|x_{S^c}\|_2 + (1 + c_T)\frac{2\|e\|_2}{\sqrt{1-\delta}} \\ &= (1 + c_T)\sqrt{\frac{1+\delta}{1-\delta}}\left\|(x - x^i)_{S^c}\right\|_2 + (1 + c_T)\frac{2\|e\|_2}{\sqrt{1-\delta}} \\ &\leq (1 + c_T)\sqrt{\frac{1+\delta}{1-\delta}}\left\|r^i_{\Gamma^c}\right\|_2 + (1 + c_T)\frac{2\|e\|_2}{\sqrt{1-\delta}} . \qquad (18)\end{aligned}$$

The first relation once again follows from the triangle inequality. The second relation follows from the fact that $\mathrm{supp}(x_{S^c}) \in \mathbb{M}^+$ (since $\mathrm{supp}(x) \in \mathbb{M}^+$), and therefore, $A_{S^c} x_{S^c}$ can be upper-bounded using the model-RIP. The third follows from the fact that x_i supported on S^c is zero because S fully subsumes the support of x^i. The final relation follows from the fact that $S^c \subseteq \Gamma^c$ (see line 6 in the algorithm).

Note that the support Γ is defined as in Lemma 9. Therefore, we can use (9) and bound $\left\|r^i_{\Gamma^c}\right\|_2$ in terms of $\left\|r^i\right\|_2$, c_H, and δ. Substituting into (18) and rearranging terms, we obtain the stated theorem. □

As in the analysis of AM-IHT, suppose that $e = 0$ and δ is very small. Then, we achieve geometric convergence, i.e., $\alpha < 1$, if the approximation factors c_T and c_H satisfy

$$(1 + c_T)\sqrt{1 - c_H^2} < 1\,, \qquad \text{or equivalently,} \qquad c_H^2 > 1 - \frac{1}{(1 + c_T)^2}\,. \tag{19}$$

Therefore, the conditions for convergence of AM-IHT and AM-CoSaMP are identical in this regime. As for AM-IHT, we relax this condition for AM-CoSaMP in Section 7 and show that geometric convergence is possible for *any* constants c_T and c_H.

6 Approximate Model-IHT with RIP-1 matrices

AM-IHT and AM-CoSaMP (Algorithms 1 and 2) rely on measurement matrices satisfying the model-RIP (Definition 3). It is known that $m \times n$ matrices whose elements are drawn i.i.d. from a sub-Gaussian distribution satisfy this property with high probability while requiring only a small number of rows m [12, 40]. However, such matrices are *dense* and consequently incur significant costs of $\Theta(m \cdot n)$ for both storage and matrix-vector multiplications.

One way to circumvent this issue is to consider *sparse* measurement matrices [6]. Sparse matrices can be stored very efficiently and enable fast matrix-vector multiplication (with both costs scaling proportionally to the number of nonzeros). However, the usual RIP does not apply for such matrices. Instead, such matrices are known to satisfy the RIP in the ℓ_1-norm (or *RIP-1*). Interestingly, it can be shown that this property is sufficient to enable robust sparse recovery for arbitrary signals [23]. Moreover, several existing algorithms for sparse recovery can be modified to work with sparse measurement matrices; see [23, 24].

In the model-based compressive sensing context, one can analogously define the RIP-1 over structured sparsity models as follows:

Definition 14 (Model RIP-1). *A matrix $A \in \mathbb{R}^{m \times n}$ has the $(\delta, \mathbb{M})$-model RIP-1 if the following holds for all x with* $\operatorname{supp}(x) \in \mathbb{M}^+$*:*

$$(1 - \delta)\|x\|_1 \leq \|Ax\|_1 \leq (1 + \delta)\|x\|_1\,. \tag{20}$$

The paper [27] establishes both lower and upper bounds on the number of measurements required to satisfy the model RIP-1 for certain structured sparsity models. Similar to Fact 6, the paper also provides a general sampling bound based on the cardinality of the model:

Fact 15 (Theorem 9 in [27]). *Let $\mathcal{M}$ be a structured sparsity model and let k be the size of the largest support in the model, i.e., $k = \max_{\Omega \in \mathbb{M}} |\Omega|$. Then there is a $m \times n$ matrix satisfying the $(\delta, \mathbb{M})$-model RIP-1 with*

$$m = O\left(\frac{k}{\delta^2} \cdot \frac{\log(n/l)}{\log(k/l)}\right),$$

where

$$l = \frac{\log|\mathbb{M}|}{\log(n/k)}\,.$$

Subsequently, the paper [25] proposes a modification of *expander iterative hard thresholding* (EIHT) [24] that achieves stable recovery for arbitrary structured sparsity models. As before, this modified algorithm only works when provided access to *exact* model-projection oracles. Below,

Algorithm 3 AM-IHT with RIP-1

```
function AM-IHT-RIP-1($y, A, t$)
    $x^0 \leftarrow 0$
    for $i \leftarrow 0, \ldots, t$ do
        $x^{i+1} \leftarrow T(x^i + H(\mathrm{MED}(y - Ax^i)))$
    return $x^{t+1}$
```

we propose a more general algorithm suitable for model-based recovery using only approximate projection oracles.

Before proceeding further, it is worthwhile to understand a particular class of matrices that satisfy the RIP-1. It is known that adjacency matrices of certain carefully chosen random bipartite graphs, known as *bipartite expanders*, satisfy the model RIP-1 [23, 27]. Indeed, suppose that such a matrix A represents the bipartite graph $G = ([n], [m], E)$, where E is the set of edges. For any $S \subseteq [n]$, define $\Gamma(S)$ to be the set of nodes in $[m]$ connected to S by an edge in E. Therefore, we can define the *median operator* $\mathrm{MED}(u) : \mathbb{R}^m \to \mathbb{R}^n$ for any $u \in \mathbb{R}^m$ component-wise as follows:

$$[\mathrm{MED}(u)]_i = \mathrm{median}[u_j : j \in \Gamma(\{i\})] \, .$$

This operator is crucial in our algorithm and proofs below.

We now propose a variant of AM-IHT (Algorithm 1) that is suitable when the measurement matrix A satisfies the RIP-1. The description of this new version is provided as Algorithm 3. Compared to AM-IHT, the important modification in the RIP-1 algorithm is the use of the median operator $\mathrm{MED}(\cdot)$ instead of the transpose of the measurement matrix A.

We analytically characterize the convergence behavior of Algorithm 3. First, we present the following Lemma, which is proved in [25] based on [24].

Lemma 16 (Lemma 7.2 in [25])**.** *Suppose that A satisfies the $(\delta, \mathbb{M})$-model-RIP-1. Then, for any vectors $x \in \mathbb{R}^n$, $e \in \mathbb{R}^m$, and any support $S \in \mathbb{M}^+$,*

$$\|[x - \mathrm{MED}(Ax_S + e)]_S\|_1 \leq \rho_0 \|x_S\|_1 + \tau_0 \|e\|_1 \, .$$

Here, $\rho_0 = 4\delta/(1 - 4\delta)$ and τ_0 is a positive scalar that depends on δ.

Armed with this Lemma, we now prove the main result of this section. We make similar assumptions as in Section 4, this time using the model-RIP-1 and approximate projection oracles for the ℓ_1-norm: (i) $x \in \mathbb{R}^n$ and $x \in \mathcal{M}$. (ii) $y = Ax + e$ for an arbitrary $e \in \mathbb{R}^m$ (the measurement noise). (iii) T is a $(c_T, \mathbb{M}, \mathbb{M}_T, 1)$-tail-approximation oracle. (iv) H is a $(c_H, \mathbb{M}_T \oplus \mathbb{M}, \mathbb{M}_H, 1)$-head-approximation-oracle. (v) A has the $(\delta, \mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H)$-model-RIP-1. Then, we obtain:

Theorem 17 (Geometric convergence of AM-IHT with RIP-1)**.** *Let $r^i = x - x^i$, where x^i is the signal estimate computed by AM-IHT-RIP-1 in iteration i. Let ρ_0, τ_0 be as defined in Lemma 16. Then, AM-IHT-RIP-1 exhibits the following convergence property:*

$$\|r^{i+1}\|_1 \leq \rho \|r^i\|_1 + \tau \|e\|_1 \, ,$$

where

$$\rho = (1 + c_T)(2\rho_0 + 1 - c_H(1 - \rho_0)) \, ,$$
$$\tau = (1 + c_T)(2 + c_H)\tau_0 \, .$$

Proof. Let $a_i = x_i + H(\mathrm{MED}(y - Ax_i))$. The triangle inequality gives:

$$\begin{aligned}\left\|r^{i+1}\right\|_1 &= \left\|x - x^{i+1}\right\|_1 \\ &\leq \left\|x - a^i\right\|_1 + \left\|x^{i+1} - a^i\right\|_1 \\ &\leq (1 + c_T)\left\|x - a^i\right\|_1 \\ &\leq (1 + c_T)\left\|x - x^i - H(\mathrm{MED}(y - Ax^i))\right\|_1 \\ &= (1 + c_T)\left\|r^i - H(\mathrm{MED}(Ar^i + e))\right\|_1 .\end{aligned}$$

Let $v = \mathrm{MED}(Ar^i + e)$, $\Omega = \mathrm{supp}(r^i)$, and Γ be the support returned by the head oracle H. We have:

$$\left\|H(v)\right\|_1 = \left\|v_\Gamma\right\|_1 \geq c_H \left\|v_\Omega\right\|_1 , \tag{21}$$

due to the head-approximation property of H.

On the other hand, we also have

$$\begin{aligned}\left\|v_\Omega - r^i\right\|_1 &= \left\|(\mathrm{MED}(Ar^i + e) - r^i)_\Omega\right\|_1 \\ &\leq \left\|(\mathrm{MED}(Ar^i + e) - r^i)_{\Omega\cup\Gamma}\right\|_1 \\ &\leq \rho_0\left\|r^i\right\|_1 + \tau_0\left\|e\right\|_1 .\end{aligned}$$

where the last inequality follows from Lemma 16 (note that we use the lemma for the model $\mathbb{M} \oplus \mathbb{M}_T \oplus \mathbb{M}_H$). Further, by applying the triangle inequality again and combining with (21), we get

$$\left\|H(v)\right\|_1 \geq c_H(1 - \rho_0)\left\|r^i\right\|_1 - c_H\tau_0\left\|e\right\|_1 . \tag{22}$$

We also have the following series of inequalities:

$$\begin{aligned}\left\|H(v)\right\|_1 &= \left\|H(v) - r^i_\Gamma + r^i_\Gamma\right\|_1 \\ &\leq \left\|v_\Gamma - r^i_\Gamma\right\|_1 + \left\|r^i_\Gamma\right\|_1 \\ &\leq \left\|v_{\Gamma\cup\Omega} - r^i_{\Gamma\cup\Omega}\right\|_1 + \left\|r^i_\Gamma\right\|_1 \\ &= \left\|(\mathrm{MED}(Ar^i + e) - r^i)_{\Omega\cup\Gamma}\right\|_1 + \left\|r^i_\Gamma\right\|_1 \\ &\leq \rho_0\left\|r^i\right\|_1 + \tau_0\left\|e\right\|_1 + \left\|r^i_\Gamma\right\|_1 .\end{aligned}$$

Here, we have once again invoked Lemma 16. Moreover, $\left\|r^i_\Gamma\right\|_1 = \left\|r^i\right\|_1 - \left\|r^i_{\Gamma^c}\right\|_1$. Combining with (22) and rearranging terms, we get:

$$\left\|r^i_{\Gamma^c}\right\|_1 \leq (\rho_0 + 1 - c_H(1 - \rho_0))\left\|r^i\right\|_1 + (1 + c_H)\tau_0\left\|e\right\|_1 . \tag{23}$$

Recall that

$$\begin{aligned}\left\|r^{i+1}\right\|_1 &\leq (1 + c_T)\left\|r^i - H(v)\right\|_1 \\ &= (1 + c_T)\left(\left\|r^i_\Gamma - v_\Gamma\right\|_1 + \left\|r^i_{\Gamma^c}\right\|_1\right) ,\end{aligned}$$

since $v_\Gamma = H(v) = H(\mathrm{MED}(Ar^i + e))$. Invoking Lemma 16 one last time and combining with (23), we obtain

$$\begin{aligned}\left\|r^{i+1}\right\|_1 &\leq (1 + c_T)\left[\rho_0\left\|r^i\right\|_1 + \tau_0\left\|e\right\|_1 + (\rho_0 + 1 - c_H(1 - \rho_0))\left\|r^i\right\|_1 + (1 + c_H)\tau_0\left\|e\right\|_1\right] \\ &\leq (1 + c_T)(2\rho_0 + 1 - c_H(1 - \rho_0))\left\|r^i\right\|_1 + (1 + c_T)(2 + c_H)\tau_0\left\|e\right\|_1 ,\end{aligned}$$

as claimed. □

Algorithm 4 Boosting for head-approximation algorithms

```
function BoostHead($x, H, t$)
    $\Omega_0 \leftarrow \{\}$
    for $i \leftarrow 1, \ldots, t$ do
        $\Lambda_i \leftarrow H(x_{[n]\setminus\Omega_{i-1}})$
        $\Omega_i \leftarrow \Omega_{i-1} \cup \Lambda_i$
    return $\Omega_t$
```

Once again, if $e = 0$ and ρ_0 is made sufficiently small, AM-IHT with RIP-1 achieves geometric convergence to the true signal x provided that $c_H > 1 - 1/(1 + c_T)$. Thus, we have developed an analogue of AM-IHT that works purely with the RIP-1 assumption on the measurement matrix and hence is suitable for recovery using sparse matrices. It is likely that a similar analogue can be developed for AM-CoSaMP, but we will not pursue this direction here.

7 Improved Recovery via Boosting

As stated in Sections 4 and 5, AM-IHT and AM-CoSaMP require stringent assumptions on the head- and tail-approximation factors c_H and c_T. The condition (16) indicates that for AM-IHT to converge, the head- and tail-approximation factors must be tightly coupled. Observe that by definition, c_T is no smaller than 1. Therefore, c_H must be at least $\sqrt{3}/2$. If c_T is large (i.e., if the tail-approximation oracle gives only a crude approximation), then the head-approximation oracle needs to be even more precise. For example, if $c_T = 10$, then $c_H > 0.995$, i.e., the head approximation oracle needs to be very accurate. Such a stringent condition can severely constrain the choice of approximation algorithms.

In this section, we overcome this barrier by demonstrating how to "boost" the approximation factor of any given head-approximation algorithm. Given a head-approximation algorithm with arbitrary approximation factor c_H, we can boost its approximation factor to any arbitrary constant $c'_H < 1$. Our approach requires only a constant number of invocations of the original head-approximation algorithm and inflates the sample complexity of the resulting output model only by a constant factor. Combining this boosted head-approximation algorithm with AM-IHT or AM-CoSaMP, we can provide an overall recovery scheme for approximation algorithms with *arbitrary* approximation constants c_T and c_H. This is a much weaker condition than (16) and therefore significantly extends the scope of our framework for model-based compressive sensing with approximate projection oracles.

We achieve this improvement by iteratively applying the head-approximation algorithm to the residual of the currently selected support. Each iteration guarantees that we add another c_H-fraction of the best remaining support to our result. Algorithm 4 contains the corresponding pseudo code and Theorem 18 the main guarantees.

Theorem 18. *Let H be a $(c_H, \mathbb{M}, \mathbb{M}_H, p)$-head-approximation algorithm with $0 < c_H \leq 1$ and $p \geq 1$. Then* BoostHead(x, H, t) *is a $((1-(1-c_H^p)^t)^{1/p}, \mathbb{M}, \mathbb{M}_H^{\oplus t}, p)$-head-approximation algorithm. Moreover,* BoostHead *runs in time $O(t \cdot T_H)$, where T_H is the time complexity of H.*

Proof. Let $\Gamma \in \mathbb{M}$ be an optimal support, i.e., $\|x_\Gamma\|_p = \max_{\Omega\in\mathbb{M}}\|x_\Omega\|_p$. We now prove that the following invariant holds at the beginning of iteration i:

$$\|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p \le (1-c_H^p)^{i-1}\|x_\Gamma\|_p^p\ . \tag{24}$$

Note that the invariant (Equation 24) is equivalent to $\left\|x_{\Omega_{i-1}}\right\|_p^p \ge \left(1-(1-c_H^p)^{i-1}\right)\|x_\Gamma\|_p^p$. For $i = t+1$, this gives the head-approximation guarantee stated in the theorem.

For $i = 1$, the invariant directly follows from the initialization.

Now assume that the invariant holds for an arbitrary $i \ge 1$. From line 4 we have

$$\begin{aligned}
\left\|(x_{[n]\setminus\Omega_{i-1}})_{\Lambda_i}\right\|_p^p &\ge c_H^p \max_{\Omega\in\mathbb{M}}\left\|(x_{[n]\setminus\Omega_{i-1}})_\Omega\right\|_p^p\\
\left\|x_{\Lambda_i\setminus\Omega_{i-1}}\right\|_p^p &\ge c_H^p \max_{\Omega\in\mathbb{M}}\left\|(x - x_{\Omega_{i-1}})_\Omega\right\|_p^p\\
&\ge c_H^p\left\|(x - x_{\Omega_{i-1}})_\Gamma\right\|_p^p\\
&= c_H^p\left\|x_\Gamma - x_{\Omega_{i-1}\cap\Gamma}\right\|_p^p\\
&= c_H^p\Big(\|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}\cap\Gamma}\right\|_p^p\Big)\\
&\ge c_H^p\Big(\|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p\Big)\ .
\end{aligned} \tag{25}$$

We now prove the invariant for $i+1$:

$$\begin{aligned}
\|x_\Gamma\|_p^p - \|x_{\Omega_i}\|_p^p &= \|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p - \left\|x_{\Lambda_i\setminus\Omega_{i-1}}\right\|_p^p\\
&\le \|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p - c_H^p\Big(\|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p\Big)\\
&= (1-c_H^p)\Big(\|x_\Gamma\|_p^p - \left\|x_{\Omega_{i-1}}\right\|_p^p\Big)\\
&\le (1-c_H^p)^{i+1}\|x_\Gamma\|_p^p\ .
\end{aligned}$$

The second line follows from (25) and the third line from the invariant.

Since $\Lambda_i \in \mathbb{M}_H$, we have $\Omega_t \in \mathbb{M}_H^{\oplus t}$. The time complexity of BoostHead follows directly from the definition of the algorithm. □

We now use Theorem 18 to relax the conditions on c_T and c_H in Corollary 12. As before, we assume that we have compressive measurements of the form $y = Ax + e$, where $x \in \mathcal{M}$ and e is arbitrary measurement noise.

Corollary 19. *Let T and H be approximate projection oracles with $c_T \ge 1$ and $0 < c_H < 1$. Moreover, let δ be the model-RIP constant of the measurement matrix A and let*

$$\begin{aligned}
\gamma &= \frac{\sqrt{1-\left(\frac{1}{1+c_T}-\delta\right)^2}+\delta}{1-\delta}\,,\\
t &= \left\lceil\frac{\log(1-\gamma^2)}{\log(1-c_H^2)}\right\rceil + 1\,.
\end{aligned}$$

We assume that δ is small enough so that $\gamma < 1$ and that A satisfies the model-RIP for $\mathbb{M}\oplus\mathbb{M}_T\oplus\mathbb{M}_H^{\oplus t}$. Then AM-IHT with T and BoostHead(x, H, t) as projection oracles returns a signal estimate $\widehat{x}$ satisfying

$$\|x-\widehat{x}\|_2 \le C\|e\|_2$$

after $O(\log \frac{\|x\|_2}{\|e\|_2})$ iterations. The constants in the error and runtime bounds depend only on c_T, c_H, and δ.

Proof. In order to use Corollary 12, we need to show that $\alpha < 1$. Recall that

$$\alpha = (1 + c_T)(\delta + \sqrt{1 - (c_H(1-\delta) - \delta)^2})\,.$$

A simple calculation shows that a head-approximation oracle with $c'_H > \gamma$ achieves $\alpha < 1$.

Theorem 18 shows that boosting the head-approximation oracle H with t' iterations gives a head-approximation factor of

$$c'_H = \sqrt{1 - (1 - c_H^2)^{t'}}\,.$$

Setting $t' = t$ as defined in the theorem yields $c'_H > \gamma$. We can now invoke Corollary 12 for the recovery guarantee of AM-IHT. □

Analogous corollaries can be proven for AM-CoSaMP (Section 5) and AM-IHT with RIP-1 (Section 6). We omit detailed statements of these results here.

8 Case Study: The CEMD model

As an instantiation of our main results, we discuss a special structured sparsity model known as the *Constrained EMD* model [26]. A key ingredient in the model is the Earth Mover's Distance (EMD), also known as the Wasserstein metric or Mallows distance [42]:

Definition 20 (EMD)**.** *The EMD of two finite sets $A, B \subset \mathbb{N}$ with $|A| = |B|$ is defined as*

$$\mathrm{EMD}(A, B) = \min_{\pi : A \to B} \sum_{a \in A} |a - \pi(a)|\,, \tag{26}$$

where π ranges over all one-to-one mappings from A to B.

Observe that $\mathrm{EMD}(A, B)$ is equal to the cost of a min-cost matching between A and B. Now, consider the case where the sets A and B are the *supports* of two exactly k-sparse signals, so that $|A| = |B| = k$. In this case, the EMD not only measures how many indices change, but also how far the supported indices move. This notion can be generalized from pairs of signals to an *ensemble* of sparse signals. Figure 1 illustrates the following definition.

Definition 21 (Support-EMD)**.** *Let $\Omega \subseteq [h] \times [w]$ be the support of a matrix X with exactly s-sparse columns, i.e., $|\text{col-supp}(\Omega, c)| = s$ for $c \in [w]$. Then the EMD of Ω is defined as*

$$\mathrm{EMD}(\Omega) = \sum_{c=1}^{w-1} \mathrm{EMD}(\text{col-supp}(\Omega, c), \text{col-supp}(\Omega, c+1))\,.$$

If the columns of X are not exactly s-sparse, we define the EMD of Ω as the minimum EMD of any support that contains Ω and has exactly s-sparse columns. Let $s = \max_{c \in [w]} |\text{col-supp}(\Omega, c)|$. Then $\mathrm{EMD}(\Omega) = \min_\Gamma \mathrm{EMD}(\Gamma)$, where $\Gamma \subseteq [h] \times [w]$, $\Omega \subseteq \Gamma$, and Γ is a support with exactly s-sparse columns, i.e., $|\text{col-supp}(\Gamma, c)| = s$ for $c \in [w]$.

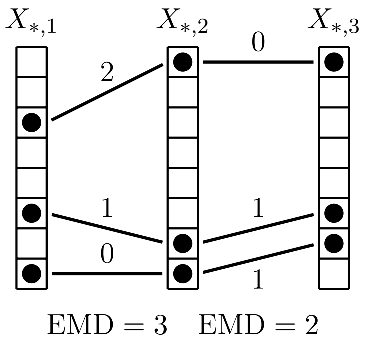

Figure (1): The support-EMD for a matrix with three columns and eight rows. The circles stand for supported elements in the columns. The lines indicate the matching between the supported elements and the corresponding EMD cost. The total support-EMD is $\mathrm{EMD}(\mathrm{supp}(X)) = 2 + 3 = 5$.

The above definitions motivate a natural structured sparsity model that essentially characterizes ensembles of sparse signals with correlated supports. Suppose we interpret the signal $x \in \mathbb{R}^n$ as a matrix $X \in \mathbb{R}^{h \times w}$ with $n = hw$. For given dimensions of the signal X, our model has two parameters: (i) k, the total sparsity of the signal. For simplicity, we assume here and in the rest of this paper that k is divisible by w. Then the sparsity of each column $X_{*,i}$ is $s = k/w$. (ii) B, the support-EMD of X. We call this parameter the *EMD budget*. Formally, we have:

Definition 22 (Constrained EMD model)**.** *The Constrained EMD (CEMD) model is the structured sparsity model $\mathcal{M}_{k,B}$ defined by the set of supports $\mathbb{M}_{k,B} = \{\Omega \subseteq [h] \times [w] \,|\, \mathrm{EMD}(\Omega) \leq B$ and $|\text{col-supp}(\Omega, c)| = \frac{k}{w}$ for $c \in [w]\}$.*

The parameter B controls how much the support can vary from one column to the next. Setting $B = 0$ forces the support to remain constant across all columns, which corresponds to block sparsity (the blocks are the rows of X). A value of $B \geq kh$ effectively removes the EMD constraint because each supported element is allowed to move across the full height of the signal. In this case, the model demands only s-sparsity in each column. It is important to note that we only constrain the EMD of the column *supports* in the signal, not the actual amplitudes. Figure 2 illustrates the CEMD model with an example.

8.1 Sampling bound

Our objective is to develop a sparse recovery scheme for the Constrained EMD model. As the first ingredient, we establish the model-RIP for $\mathcal{M}_{k,B}$, i.e., we characterize the number of permissible supports (or equivalently, the number of subspaces) $l_{k,B}$ in the model and invoke Fact 6. For simplicity, we will assume that $w = \Omega(\log h)$, i.e., the following bounds apply for all signals X except very thin and tall matrices X. The following result is novel:

Theorem 23. *The number of allowed supports in the CEMD model satisfies $\log|\mathbb{M}_{k,B}| = O\big(k \log \frac{B}{k}\big)$.*

Proof. For given h, w, B, and k, the support is fixed by the following three decisions: (i) The choice of the supported elements in the first column of X. (ii) The distribution of the EMD budget B over the k supported elements. This corresponds to distributing B balls into $k+1$ bins (using one bin for

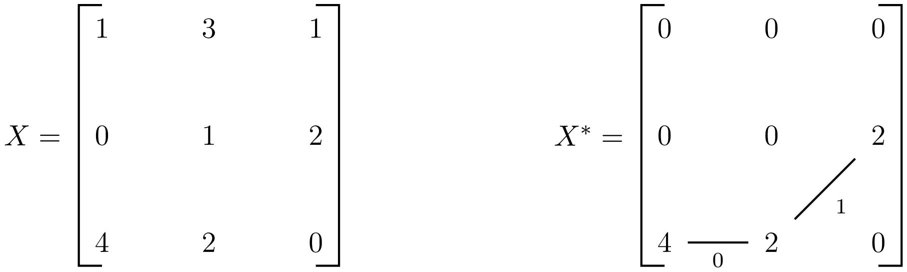

Figure (2): A signal X and its best approximation X^* in the EMD model $\mathcal{M}_{3,1}$. A sparsity constraint of 3 with 3 columns implies that each column has to be 1-sparse. Moreover, the total support-EMD between neighboring columns in X^* is 1. The lines in X^* indicate the support-EMD.

the part of the EMD budget not allocated to supported elements). (iii) For each supported element, the direction (up or down) to the matching element in the next column to the right. Multiplying the choices above gives $\binom{h}{s}\binom{B+k}{k}2^k$, an upper bound on the number of supports. Using the inequality $\binom{a}{b} \le \left(\frac{a\,e}{b}\right)^b$, we get

$$\begin{aligned}\log|\mathbb{M}_{k,B}| &\le \log\left(\binom{h}{s}\binom{B+k}{k}2^k\right)\\ &\le s\log\frac{h}{s} + k\log\frac{B+k}{k} + O(s+k)\\ &= O\left(k\log\frac{B}{k}\right).\end{aligned}$$ □

If we allow each supported element to move a constant amount from one column to the next, we get $B = O(k)$ and hence, from Fact 6, $m = O(k + \log|\mathbb{M}_{k,B}|) = O(k)$ rows for sub-Gaussian measurement matrices. This bound is information-theoretically optimal. Furthermore, for $B = kh$ (i.e., allowing every supported element to move anywhere in the next column) we get $m = O(k\log n)$, which almost matches the standard compressive sensing bound of $m = O(k\log\frac{n}{k})$ for sub-Gaussian measurement matrices. Therefore, the CEMD model gives a smooth trade-off between the support variability and the number of measurements necessary for recovery.

We can also establish a sampling bound in the RIP-1 setting with Fact 15. For the case of $B = \Theta(k)$, we get $m = O(k\frac{\log n}{\log\log\frac{n}{k}})$. In order to match the block-sparsity lower bound of $m = O(k\log_w n)$, we need to assume that $B = O(k/w)$, i.e., each path (and not each element) in the support has a constant EMD-budget on average. We omit the details of this calculation here.

The following theorem is useful when establishing sampling bounds for recovery schemes using the CEMD model.

Theorem 24. *The CEMD model is closed under addition:* $\mathbb{M}_{k_1,B_1} \oplus \mathbb{M}_{k_2,B_2} \subseteq \mathbb{M}^+_{k_1+k_2,B_1+B_2}$.

Proof. Let $\Omega_1 \in \mathbb{M}_{k_1,B_1}$ and $\Omega_2 \in \mathbb{M}_{k_2,B_2}$. Moreover, let $\Gamma = \Omega_1 \cup \Omega_2$. We have to show that $\Gamma \in \mathbb{M}_{k_1+k_2,B_1+B_2}$.

The column-sparsity of Ω_1 and Ω_2 is k_1/w and k_2/w, respectively. Hence the column-sparsity of Γ is at most $\frac{k_1+k_2}{w}$. Moreover, we can construct a matching for Γ with cost at most $B_1 + B_2$ from the matchings for Ω_1 and Ω_2. To see this, consider without loss of generality the matchings π_1 and

π_2 corresponding to the first two columns in Ω_1 and Ω_2, respectively. We start constructing the new matching π' by starting with π_1. Then, we iterate over the pairs (a, b) in π_2 one by one and augment π' to include both a and b. There are four cases:

1. Both a and b are still unassigned in π'. Then we can simply add (a, b) to π'.
2. Both a and b are already assigned in π'. In this case, we do not need to modify π' to include a and b.
3. a is not included in π', but b is already assigned in π'. This is the interesting case becaues we must now find a new neighbor assignment for a. Let b' be the entry in the second column that is in the same row as a. If b' is not assigned yet, we can simply add (a, b') to π'. Otherwise, let a' be the value such that $\pi'(a') = b'$. Then we remove the pair (a', b') from π', add (a, b') to π', and repeat this procedure to find a new neighbor for a'. It is easy to see that this procedure terminates after a finite number of steps, and that no node currently assigned under π' loses a neighbor. Moreover, note that this operation does not increase the cost of the matching π'.
4. b is not included in π', but a is already assigned in π'. This case is symmetric to case 3 above.

Each of the four cases increases the cost of π' by at most the cost of (a, b) in π_2. Iterating over all pairs in π_2, we observe that the final matching π' has cost no more than the cumulative costs of π_1 and π_2, i.e., at most $B_1 + B_2$. Therefore, $\Gamma \in \mathbb{M}_{k_1+k_2,B_1+B_2}$. □

8.2 Head Approximation Algorithm

First, we develop a head approximation algorithm for the CEMD model. Ideally, we would have an *exact* projection algorithm H mapping arbitrary signals to signals in $\mathcal{M}_{k,B}$ with the guarantee $\|H(x)\|_p = \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_p$. However, this appears to be a hard problem. Instead, we propose an efficient greedy algorithm satisfying the somewhat looser requirements of a head approximation oracle (Definition 7). Specifically, we develop an algorithm that performs the following task: given an arbitrary signal x, find a support $\Omega \in \mathbb{M}_{O(k),O(B \log k)}$ such that $\|x_\Omega\|_p^p \geq c \max_{\Gamma \in \mathbb{M}_{k,B}} \|x_\Gamma\|_p^p$, where $c > 0$ is a fixed constant.

As before, we interpret our signal x as a matrix $X \in \mathbb{R}^{h \times w}$. Let OPT denote the largest sum of coefficients achievable with a support in $\mathbb{M}_{k,B}$, i.e., $OPT = \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_p^p$. For a signal $x \in \mathcal{M}_{k,B}$, we interpret the support of x as a set of $s = k/w$ paths from the leftmost to the rightmost column in X. Our method proceeds by greedily finding a set of paths that cover a large sum of signal coefficients. We can then show that the coefficients covered by these paths are a constant fraction of the optimal coefficient sum OPT.

Definition 25 (Path in a matrix). *Given a matrix $X \in \mathbb{R}^{h \times w}$, a path $r \subseteq [h] \times [w]$ is a set of w locations in X with one location per column, i.e., $|r| = w$ and $\bigcup_{(i,j) \in r} j = [w]$. The weight of r is the sum of amplitudes on r, i.e., $w_{X,p}(r) = \sum_{(i,j) \in r} |X_{i,j}|^p$. The EMD of r is the sum of the EMDs between locations in neighboring columns. Let $j_1, \ldots, j_w$ be the locations of r in columns 1 to w. Then,* $\mathrm{EMD}(r) = \sum_{i=1}^{w-1} |j_i - j_{i+1}|$.

Trivially, we have that a path r in X is a support with $w_{X,p}(r) = \|X_r\|_p^p$ and $\mathrm{EMD}(r) = \mathrm{EMD}(\mathrm{supp}(X_r))$. Therefore, we can iteratively build a support Ω by finding s paths in X. Algorithm 5 contains the description of HeadApprox. We show that HeadApprox finds a constant

Algorithm 5 Head approximation algorithm

```
function HeadApprox($X, k, B$)
    $X^{(1)} \leftarrow X$
    for $i \leftarrow 1, \ldots, s$ do
        Find the path $r_i$ from column 1 to column $w$ in $X^{(i)}$ that maximizes $w^{(i)}(r_i)$ and
            uses at most EMD-budget $\lfloor \frac{B}{i} \rfloor$.
        $X^{(i+1)} \leftarrow X^{(i)}$
        for $(u, v) \in r_i$ do
            $X^{(i+1)}_{u,v} \leftarrow 0$
    return $\bigcup_{i=1}^{s} r_i$
```

fraction of the amplitude sum of the best support while only moderately increasing the size of the model. For simplicity, denote $w(r) := w_{X,p}(r)$, and $w^{(i)}(r) := w_{X^{(i)},p}(r)$. We obtain the following result:

Theorem 26. *Let $p \geq 1$ and $B' = \lceil H_s \rceil B$, where $H_s = \sum_{i=1}^{s} 1/i$ is the s-th harmonic number. Then* HeadApprox *is a $((\frac{1}{4})^{1/p}, \mathbb{M}_{k,B}, \mathbb{M}_{k,B'}, p)$-head-approximation oracle.*

Proof. Let Ω be the support returned by HeadApprox(X, k, B) and let $\Omega_{OPT} \in \mathbb{M}_{k,B}$ be an optimal support. We can always decompose Ω_{OPT} into s disjoint paths in X. Let $t_1, \ldots, t_s$ be such a decomposition with $\text{EMD}(t_1) \geq \text{EMD}(t_2) \geq \ldots \geq \text{EMD}(t_s)$. Note that $\text{EMD}(t_i) \leq \lfloor \frac{B}{i} \rfloor$: otherwise $\sum_{j=1}^{i} \text{EMD}(t_i) > B$ and since $\text{EMD}(\Omega_{OPT}) \leq B$ this would be a contradiction. Since Ω is the union of s disjoint paths in X, Ω has column-sparsity s. Moreover, we have $\text{EMD}(\Omega) = \sum_{i=1}^{s} \text{EMD}(r_i) \leq \sum_{i=1}^{s} \lfloor \frac{B}{i} \rfloor \leq \lceil H_s \rceil B$. Therefore, $\Omega \in \mathbb{M}^{+}_{k,B'}$.

When finding path r_i in $X^{(i)}$, there are two cases:

Case 1: $w^{(i)}(t_i) \leq \frac{1}{2} w(t_i)$, i.e., the paths $r_1, \ldots, r_{i-1}$ have already covered more than half of the coefficient sum of t_i in X.

Case 2: $w^{(i)}(t_i) > \frac{1}{2} w(t_i)$, i.e., there is still more than half of the coefficient sum of t_i remaining in $X^{(i)}$. Since $\text{EMD}(t_i) \leq \lfloor \frac{B}{i} \rfloor$, the path t_i is a candidate when searching for the optimal path r_i and hence we find a path r_i with $w^{(i)}(r_i) > \frac{1}{2} w(t_i)$.

Let $C = \{i \in [s] \mid \text{case 1 holds for } r_i\}$ and $D = \{i \in [s] \mid \text{case 2 holds for } r_i\}$ (note that $C = [s] \setminus D$). Then we have

$$\begin{aligned} \|X_\Omega\|_p^p = \sum_{i=1}^{s} w^{(i)}(r_i) &= \sum_{i \in C} w^{(i)}(r_i) + \sum_{i \in D} w^{(i)}(r_i) \\ &\geq \sum_{i \in D} w^{(i)}(r_i) \geq \frac{1}{2} \sum_{i \in D} w(t_i) \,. \end{aligned} \tag{27}$$

For each t_i with $i \in C$, let $E_i = t_i \cap \bigcup_{j<i} r_j$, i.e., the locations of t_i already covered by some r_j when searching for r_i. Then we have

$$\sum_{(u,v) \in E_i} |X_{u,v}|^p = w(t_i) - w^{(i)}(t_i) \geq \frac{1}{2} w(t_i) \,,$$

and

$$\sum_{i \in C} \sum_{(u,v) \in E_i} |X_{u,v}|^p \geq \frac{1}{2} \sum_{i \in C} w(t_i) \,.$$

The t_i are pairwise disjoint, and so are the E_i. For every $i \in C$ we have $E_i \subseteq \bigcup_{j=1}^{s} r_j$. Hence

$$\|X_\Omega\|_p^p = \sum_{i=1}^{s} w^{(i)}(r_i) \geq \sum_{i \in C} \sum_{(u,v) \in E_i} |X_{u,v}|^p \geq \frac{1}{2} \sum_{i \in C} w(t_i) \,. \tag{28}$$

Combining Equations 27 and 28 gives:

$$\begin{aligned} 2\|X_\Omega\|_p^p \; &\geq \frac{1}{2} \sum_{i \in C} w(t_i) + \frac{1}{2} \sum_{i \in D} w(t_i) = \frac{1}{2} OPT \\ \|X_\Omega\|_p \; &\geq \left(\frac{1}{4}\right)^{1/p} \max_{\Omega' \in \mathbb{M}_{k,B}} \|X_{\Omega'}\|_p \,. \end{aligned}$$

□

Theorem 27. HeadApprox *runs in* $O(snBh)$ *time.*

Proof. Observe that the running time of HeadApprox depends on the running time of finding a path with maximum weight for a given EMD budget. The search for such a path can be performed by *dynamic programming* over a graph with $whB = nB$ nodes, or equivalently "states" of the dynamic program.[2] Each state in the graph corresponds to a state in the dynamic program, i.e., a location $(i, j) \in [w] \times [h]$ and the current amount of EMD already used $b \in \{0, 1, \ldots, B\}$. At each state, we store the largest weight achieved by a path ending at the corresponding location (i, j) and using the corresponding amount of EMD budget b. Each state has h outgoing edges to the states in the next column (given the current location, the decision on the next location also fixes the new EMD amount). Hence the time complexity of finding one largest-weight path is $O(nBh)$ (the state space has size $O(nB)$ and each update requires $O(h)$ time). Since we repeat this procedure s times, the overall time complexity of HeadApprox is $O(snBh)$. □

We can achieve an arbitrary constant head-approximation ratio by combining HeadApprox with BoostHead (see Section 7). The resulting algorithm has the same time complexity as HeadApprox. Moreover, the sparsity and EMD budget of the resulting support is only a constant factor larger than k and B'.

8.3 Tail-Approximation Algorithm

Next, we develop a tail-approximation algorithm for the CEMD model. Given an arbitrary signal x, our objective is to find a support $\Gamma \in \mathbb{M}_{k,O(B)}$ such that

$$\|x - x_\Gamma\|_p \leq c \min_{\Omega \in \mathbb{M}_{k,B}} \|x - x_\Omega\|_p \,, \tag{29}$$

where c is a constant. Note that we allow a constant factor increase in the EMD budget of the result. The algorithm we develop is precisely the graph-based approach initially proposed in [26];

[2] We use the terminology "states" here to distinguish the dynamic program from the graph we will introduce in Section 8.3.

however, our analysis here is rigorous and novel. Two core elements of the algorithm are the notions of a *flow network* and the *min-cost max-flow problem*, which we now briefly review. We refer the reader to [43] for an introduction to the graph-theoretic definitions and algorithms we employ.

The min-cost max-flow problem is a generalization of the classical maximum flow problem [44, 43]. In this problem, the input is a graph $G = (V, E)$ with designated source and sink nodes in which every edge has a certain capacity. The goal is to find an assignment of flow to edges such that the total flow from source to sink is maximized. The flow must also be valid, i.e., the amount of flow entering any intermediate node must be equal to the amount of flow leaving that intermediate node, and the amount of flow on any edge can be at most the capacity of that edge.

In the min-cost max-flow problem, every edge e also has a cost c_e (in addition to the capacity as before). The goal now is to find a flow $f : E \to \mathbb{R}_0^+$ with maximum capacity such that the cost of the flow, i.e., $\sum_{e \in E} c_e \cdot f(e)$, is minimized. One important property of the min-cost max-flow problem is that it still admits *integral* solutions if the edge capacities are integer.

Fact 28 (Theorem 9.10 in [43])**.** *If all edge capacities, the source supply, and the sink demand are integers, then there is always an integer min-cost max-flow.*

The min-cost max-flow problem has many applications, and several efficient algorithms are known [43]. We leverage this problem for our tail-approximation task by carefully constructing a suitable flow network, which we now define.

Definition 29 (EMD flow network)**.** *For a given signal X, sparsity k, and a parameter $\lambda > 0$, the flow network $G_{X,k,\lambda}$ consists of the following elements:*

- *The* nodes *comprise a source, a sink and a node $v_{i,j}$ for $i \in [h]$, $j \in [w]$, i.e., one node per entry in X (besides source and sink).*
- *G has an* edge *from every $v_{i,j}$ to every $v_{k,j+1}$ for $i, k \in [h]$, $j \in [w-1]$. Moreover, there is an edge from the source to every $v_{i,1}$ and from every $v_{i,w}$ to the sink.*
- *The* capacity *on every edge and node (except source and sink) is 1.*
- *The* cost *of node $v_{i,j}$ is $-|X_{i,j}|^p$. The cost of an edge from $v_{i,j}$ to $v_{k,j+1}$ is $\lambda|i-k|$. The cost of the source, the sink, and all edges incident to the source or sink is 0.*
- *The* supply *at the source is s $(= \frac{k}{w})$ and the demand at the sink is s.*

Figure 3 illustrates this definition with an example. The main idea is that a set of disjoint paths through the network $G_{X,k,\lambda}$ corresponds to a support in X. For any fixed value of λ, a solution of the min-cost max-flow problem on the flow network reveals a subset S of the nodes that corresponds to a support with exactly s indices per column and minimizes $-\|X_\Omega\|_p^p + \lambda \mathrm{EMD}(\Omega)$ for different choices of support Ω. In other words, the min-cost flow solves a *Lagrangian relaxation* of the original problem (29). See Lemmas 31 and 32 for a more formal statement of this connection.

A crucial issue is the choice of the Lagrange parameter λ, which defines a trade-off between the size of the tail approximation error and the support-EMD. Note that the optimal support Ω with parameters k and B does not necessarily correspond to *any* setting of λ. Nevertheless, we show that the set of supports we explore by varying λ contains a sufficiently good approximation: the tail error and the parameters k and B are only increased by constant factors compared to the optimal support Ω. Moreover, we show that we can find such a good support efficiently via a binary search over λ. Before stating our algorithm and the main result, we formalize the connection between flows and supports.

$$X = \begin{bmatrix} 1 & 3 \\ 0 & -1 \\ 2 & 1 \end{bmatrix} \qquad G_{X,k,\lambda} =$$

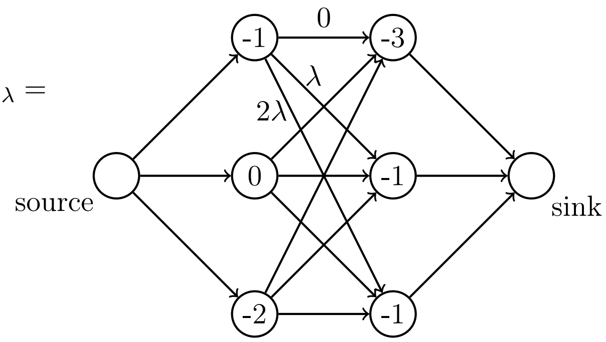

Figure (3): A signal X with the corresponding flow network $G_{X,k,\lambda}$ for $p = 1$. The node costs are the negative absolute values of the corresponding signal components. The numbers on edges indicate the edge costs (most edge costs are omitted for clarity). All capacities in the flow network are 1. The edge costs are the vertical distances between the start and end nodes, multiplied by λ.

Definition 30 (Support of a set of paths)**.** *Let $X \in \mathbb{R}^{h\times w}$ be a signal matrix, k be a sparsity parameter, and $\lambda \geq 0$. Let $P = \{q_1, \ldots, q_s\}$ be a set of disjoint paths from source to sink in $G_{X,k,\lambda}$ such that no two paths in P intersect vertically (i.e., if the q_i are sorted vertically and $i \leq j$, then $(u, v) \in q_i$ and $(w, v) \in q_j$ implies $u < w$). Then the paths in P define a support*

$$\Omega_P = \{(u, v) \mid (u, v) \in q_i \text{ for some } i \in [s]\} . \tag{30}$$

Lemma 31. *Let $X \in \mathbb{R}^{h\times w}$ be a signal matrix, k be a sparsity parameter and $\lambda \geq 0$. Let $P = \{q_1, \ldots, q_s\}$ be a set of disjoint paths from source to sink in $G_{X,k,\lambda}$ such that no two paths in P intersect vertically. Finally, let f_P be the flow induced in $G_{X,k,\lambda}$ by sending a single unit of flow along each path in P and let $c(f_P)$ be the cost of f_P. Then*

$$c(f_P) = -\|X_{\Omega_P}\|_p^p + \lambda \,\mathrm{EMD}(\Omega_P) . \tag{31}$$

Proof. The theorem follows directly from the definition of $G_{X,k,\lambda}$ and Ω_P. The node costs of P result in the term $-\|X_{\Omega_P}\|_p^p$. Since the paths in P do not intersect vertically, they are a min-cost matching for the elements in Ω_P. Hence the cost of edges between columns of X sums up to $\lambda \,\mathrm{EMD}(\Omega_P)$. □

For a fixed value of λ, a min-cost flow in $G_{X,k,\lambda}$ gives an optimal solution to the Lagrangian relaxation:

Lemma 32. *Let $G_{X,k,\lambda}$ be an EMD flow network and let f be an integral min-cost flow in $G_{X,k,\lambda}$. Then f can be decomposed into s disjoint paths $P = \{q_1, \ldots, q_s\}$ which do not intersect vertically. Moreover,*

$$\|X - X_{\Omega_P}\|_p^p + \lambda\mathrm{EMD}(\Omega_P) = \min_{\Omega \in \mathbb{M}_{k,B}} \|X - X_\Omega\|_p^p + \lambda\mathrm{EMD}(\Omega) . \tag{32}$$

Proof. Note that $\|X - X_\Omega\|_p^p = \|X\|_p^p - \|X_\Omega\|_p^p$. Since $\|X\|_p^p$ does not depend on Ω, minimizing $\|X - X_\Omega\|_p^p + \lambda\mathrm{EMD}(\Omega)$ with respect to Ω is equivalent to minimizing $-\|X_\Omega\|_p^p + \lambda\mathrm{EMD}(\Omega)$.

Further, all edges and nodes in $G_{X,k,\lambda}$ have capacity one, so f can be composed into exactly s disjoint paths P. Moreover, the paths in P are not intersecting vertically: if q_i and q_j intersect vertically, we can relax the intersection to get a set of paths P' with smaller support EMD and

Algorithm 6 Tail approximation algorithm

```
function TailApprox($X, k, B, d, \delta$)
    $x_{\min} \leftarrow \min_{|X_{i,j}|>0} |X_{i,j}|^p$
    $\varepsilon \leftarrow \frac{x_{\min}}{wh^2}\delta$
    $\lambda_0 \leftarrow \frac{x_{\min}}{2wh^2}$
    $\Omega \leftarrow$ MinCostFlow($G_{X,k,\lambda_0}$)
    if $\Omega \in \mathbb{M}_{k,B}$ and $\|X - X_\Omega\|_p = 0$ then
        return $\Omega$
    $\lambda_r \leftarrow 0$
    $\lambda_l \leftarrow \|X\|_p^p$
    while $\lambda_l - \lambda_r > \varepsilon$ do
        $\lambda_m \leftarrow (\lambda_l + \lambda_r)/2$
        $\Omega \leftarrow$ MinCostFlow($G_{X,k,\lambda_m}$)
        if EMD($\Omega$) $\geq B$ and EMD($\Omega$) $\leq dB$ then
            return $\Omega$
        if EMD($\Omega$) $> B$ then
            $\lambda_r \leftarrow \lambda_m$
        else
            $\lambda_l \leftarrow \lambda_m$
    $\Omega \leftarrow$ MinCostFlow($G_{X,k,\lambda_l}$)
    return $\Omega$
```

hence a flow with smaller cost – a contradiction. Moreover, each support $\Omega \in \mathbb{M}_{k,B}$ gives rise to a set of disjoint, not vertically intersecting paths Q and thus also to a flow f_Q with $c(f_Q) = -\|X_{\Omega_Q}\|_p^p + \lambda \mathrm{EMD}(\Omega_Q)$. Since f is a min-cost flow, we have $c(f) \leq c(f_Q)$. The statement of the theorem follows. □

We can now state our tail-approximation algorithm TailApprox (see Algorithm 6). The parameters d and δ for TailApprox quantify the acceptable tail approximation ratio (see Theorem 34). In the algorithm, we assume that MinCostFlow($G_{X,k,\lambda}$) returns the support corresponding to an integral min-cost flow in $G_{X,k,\lambda}$. Before we prove the main result (Theorem 34), we show that TailApprox always returns an optimal result for signals $X \in \mathcal{M}_{k,B}$.

Lemma 33. *Let* $x_{\min} = \min_{|X_{i,j}|>0} |X_{i,j}|^p$ *and* $\lambda_0 = \frac{x_{\min}}{2wh^2}$*. Moreover, let* $X \in \mathcal{M}_{k,B}$ *and* Ω *be the support returned by* MinCostFlow(G_{X,k,λ_0})*. Then* $\|X - X_\Omega\|_p = 0$ *and* $\Omega \in \mathbb{M}_{k,B}^+$*.*

Proof. Let $\Gamma = \mathrm{supp}(X)$, so $\Gamma \in \mathbb{M}_{k,B}^+$. First, we show that $\|X - X_\Omega\|_p = 0$. For contradiction, assume that $\|X - X_\Omega\|_p^p > 0$, so $\|X - X_\Omega\|_p^p \geq x_{\min} > 0$ (tail-approximation is trivial for $X = 0$). Since Ω is a min-cost flow, Lemma 32 gives

$$\begin{aligned} x_{\min} \; \leq \; \|X - X_\Omega\|_p^p + \lambda_0 \mathrm{EMD}(\Omega) &= \min_{\Omega' \in \mathbb{M}_{k,B}} \|X - X_{\Omega'}\|_p^p + \lambda_0 \mathrm{EMD}(\Omega') \\ &\leq 0 + \frac{x_{\min}}{2wh^2} \mathrm{EMD}(\Gamma) \\ &\leq \frac{x_{\min}}{2}\,, \end{aligned}$$

which gives a contradiction. The last line follows from $\mathrm{EMD}(\Gamma) \leq kh \leq nh$.

Now, we show that $\Omega \in \mathbb{M}^{+}_{k,B}$. By construction of G_{X,k,λ_0}, Ω is s-sparse in each column. Moreover,

$$\begin{aligned}\|X - X_\Omega\|_p^p + \lambda_0 \mathrm{EMD}(\Omega) &= \min_{\Omega' \in \mathbb{M}_{k,B}} \|X - X_\Omega\|_p^p + \lambda_0 \mathrm{EMD}(\Omega') \\ \lambda_0 \mathrm{EMD}(\Omega) &\leq 0 + \lambda_0 \mathrm{EMD}(\Gamma)\,.\end{aligned}$$

So $\mathrm{EMD}(\Omega) \leq \mathrm{EMD}(\Gamma) \leq B$. □

Next, we prove a *bicriterion*-approximation guarantee for TailApprox that allows us to use TailApprox as a tail approximation algorithm. In particular, we show that one of the following two cases occurs:

Case 1: The tail-approximation error achieved by our solution is at least as good as the best tail-approximation error achievable with support-EMD B. The support-EMD of our solution is at most a constant times larger than B.

Case 2: Our solution has bounded tail-approximation error and support-EMD at most B.

In order to simplify the proof of the main theorem, we use the following shorthands: $\Omega_l = \textsc{MinCostFlow}(G_{X,k,\lambda_l})$, $\Omega_r = \textsc{MinCostFlow}(G_{X,k,\lambda_r})$, $b_l = \mathrm{EMD}(\Omega_l)$, $b_r = \mathrm{EMD}(\Omega_r)$, $t_l = \|X - X_{\Omega_l}\|_p^p$, and $t_r = \|X - X_{\Omega_r}\|_p^p$.

Theorem 34. *Let $d > 1$, $\delta > 0$, and let Ω be the support returned by* TailApprox*(X, k, B, d, δ). Let OPT be the tail approximation error of the best support with support-EMD at most B, i.e., $OPT = \min_{\Gamma \in \mathbb{M}_{k,B}} \|X - X_\Gamma\|_p^p$. Then at least one of the following two guarantees holds for Ω:*

Case 1: $B \leq \mathrm{EMD}(\Omega) \leq dB$ *and* $\|X - X_\Omega\|_p^p \leq OPT$

Case 2: $\mathrm{EMD}(\Omega) \leq B$ *and* $\|X - X_\Omega\|_p^p \leq (1 + \frac{1}{d-1} + \delta) OPT$.

Proof. We consider the three cases in which TailApprox returns a support. If TailApprox returns in line 7, the first guarantee in the theorem is satisfied. If TailApprox reaches the binary search (line 10), we have $X \notin \mathcal{M}_{k,B}$ (the contrapositive of Lemma 33). Therefore, we have $OPT \geq x_{\min} > 0$ in the remaining two cases.

If TailApprox returns in line 14, we have $B \leq \mathrm{EMD}(\Omega) \leq dB$. Moreover, Lemma 32 gives

$$\begin{aligned}\|X - X_\Omega\|_p^p + \lambda_m \mathrm{EMD}(\Omega) &\leq \min_{\Omega' \in \mathbb{M}_{k,B}} \|X - X_{\Omega'}\|_p^p + \lambda_m \mathrm{EMD}(\Omega') \\ &\leq OPT + \lambda_m B\,.\end{aligned}$$

Since $\mathrm{EMD}(\Omega) \geq B$, we have $\|X - X_\Omega\|_p^p \leq OPT$.

We now consider the third return statement (line 20), in which case the binary search terminated with $\lambda_l - \lambda_r \leq \varepsilon$. In the binary search, we maintain the invariant that $b_l \leq B$ and $b_r > dB$. Note that this is true before the first iteration of the binary search due to our initial choices of λ_r and λ_l.[3] Moreover, our update rule maintains the invariant.

[3] Intuitively, our initial choices make the support-EMD very cheap and very expensive compared to the tail approximation error.

We now prove the bound on $\|X - X_\Omega\|_p^p = t_l$. From Lemma 32 we have

$$\begin{aligned} t_r + \lambda_r b_r &\leq OPT + \lambda_r B \\ \lambda_r dB &\leq OPT + \lambda_r B \\ \lambda_r &\leq \frac{OPT}{B(d-1)} \,. \end{aligned}$$

Since the binary search terminated, we have $\lambda_l \leq \lambda_r + \varepsilon$. We now combine this inequality with our new bound on λ_r and use it in the following inequality (also from Lemma 32):

$$\begin{aligned} t_l + \lambda_l b_l &\leq OPT + \lambda_l B \\ t_l &\leq OPT + \lambda_l B \\ &\leq OPT + (\lambda_r + \varepsilon)B \\ &\leq OPT + \frac{OPT}{d-1} + \varepsilon B \\ &\leq \left(1 + \frac{1}{d-1}\right) OPT + \frac{x_{\min}\delta B}{wh^2} \\ &\leq \left(1 + \frac{1}{d-1}\right) OPT + \delta x_{\min} \\ &\leq \left(1 + \frac{1}{d-1} + \delta\right) OPT \,. \end{aligned}$$

This shows that the second guarantee of the theorem is satisfied. □

Corollary 35. *Let $p \geq 1$, $c > 1$, $0 < \delta < c - 1$, and $d = 1 + \frac{1}{c-\delta-1}$. Then* TailApprox *is a $(c^{1/p}, \mathbb{M}_{k,B}, \mathbb{M}_{k,dB}, p)$-tail approximation algorithm.*

Proof. The tail approximation guarantee follows directly from Theorem 34. Note that we cannot control which of the two guarantees the algorithm returns. However, in any case we have $\mathrm{EMD}(\Omega) \leq dB$, so $\Omega \in \mathbb{M}_{k,dB}$. □

In order to simplify the time complexity of TailApprox, we assume that $h = \Omega(\log w)$, i.e., the matrix X is not very "wide" and "short". We arrive at the following result.

Theorem 36. *Let $\delta > 0$, $x_{\min} = \min_{|X_{i,j}|>0} |X_{i,j}|^p$, and $x_{\max} = \max |X_{i,j}|^p$. Then* TailApprox *runs in $O(snh(\log \frac{n}{\delta} + \log \frac{x_{\max}}{x_{\min}}))$ time.*

Proof. We can solve our instances of the min-cost flow problem by finding s augmenting paths because all edges and nodes have unit capacity. Moreover, $G_{X,k,\lambda}$ is a directed acyclic graph, so we can compute the initial node potentials in linear time. Each augmenting path can then be found with a single run of Dijkstra's algorithm, which costs $O(wh \log(wh) + wh^2) = O(nh)$ time [44]. The number of iterations of the binary search is at most

$$\log \frac{\|X\|_p^p}{\epsilon} \quad = \quad \log \frac{\|X\|_p^p nh}{x_{\min}\delta} \quad \leq \quad \log \frac{x_{\max} n^2 h}{x_{\min}\delta} \quad \leq \quad \log \frac{n^3}{\delta} + \log \frac{x_{\max}}{x_{\min}} \,. \tag{33}$$

Combining this with a per-iteration cost of $O(snh)$ gives the stated running time. □

To summarize, the algorithm proposed in [26] satisfies the criteria of a tail-approximation oracle. This, in conjunction with the head approximation oracle proposed in Section 8.2, gives a full sparse recovery scheme for the CEMD model, which we describe below.

8.4 Compressive Sensing Recovery

We now bring the results from the previous sections together. Specifically, we show that AM-IHT (Algorithm 1), equipped with HeadApprox and TailApprox, constitutes a model-based compressive sensing recovery algorithm that significantly reduces the number of measurements necessary for recovering signals in the CEMD model. The main result is the following theoretical guarantee:

Theorem 37. *Let $x \in \mathcal{M}_{k,B}$ be an arbitrary signal in the CEMD model with dimension $n = wh$. Let $A \in \mathbb{R}^{m\times n}$ be a measurement matrix with i.i.d. Gaussian entries and let $y \in \mathbb{R}^m$ be a noisy measurement vector, i.e., $y = Ax+e$ with arbitrary $e \in \mathbb{R}^m$. Then we can recover a signal approximation $\widehat{x} \in \mathcal{M}_{k,2B}$ satisfying $\|x - \widehat{x}\|_2 \leq C\|e\|_2$ for some constant C from $m = O(k\log(\frac{B}{k}\log\frac{k}{w}))$ measurements. Moreover, the recovery algorithm runs in time $O(n\log\frac{\|x\|_2}{\|e\|_2}(k\log n+\frac{kh}{w}(B+\log n+\log\frac{x_{\max}}{x_{\min}})))$ where $x_{\min} = \min_{|x_i|>0}|x_i|$ and $x_{\max} = \max|x_i|$.*

Proof. First, we show that m rows suffice for A to have the desired model-RIP. Following the conditions in Corollary 19, A must satisfy the $(\delta, \mathbb{M}_{k,B} \oplus \mathbb{M}_T \oplus \mathbb{M}_H^{\oplus t})$-model-RIP for small δ, where t is the number of times we boost HeadApprox (a constant depending on δ and c_T). We have $\mathbb{M}_T = \mathbb{M}_{k,2B}$ from Corollary 35 and $\mathbb{M}_H = \mathbb{M}_{2k,3\gamma B}$ where $\gamma = \lceil\log\frac{k}{w}\rceil + 1$ from Theorems 24 and 26 (note that HeadApprox must be a $(c_H, \mathbb{M}\oplus\mathbb{M}_T, \mathbb{M}_H, 2)$-head-approximation oracle). Invoking Theorem 24 again shows that it suffices for A to have the $(\delta, \mathbb{M}_{(2+2t)k,(3+3t\gamma)B})$-model-RIP. Using Theorem 23 and the fact that t is a constant, Fact 6 then shows that

$$m \;=\; O\left(k\log\frac{\gamma B}{k}\right) \;=\; O\left(k\log\left(\frac{B}{k}\log\frac{k}{w}\right)\right)$$

suffices for A to have the desired model-RIP.

Equipped with our model-RIP, we are now able to invoke Corollary 19, which directly gives the desired recovery guarantee $\|x - \widehat{x}\|_2 \leq C\|e\|_2$. Moreover, the corollary also shows that the number of iterations of AM-IHT is bounded by $O(\log\frac{\|x\|_2}{\|e\|_2})$. In order to prove our desired time complexity, we now only have to bound the per-iteration cost of AM-IHT.

In each iteration of AM-IHT, the following operations have a relevant time complexity: (i) Multiplication with A and A^T. The measurement matrix has at most $k\log n$ rows, so we bound this time complexity by $O(nk\log n)$. (ii) HeadApprox. From Theorem 27 we know that HeadApprox runs in time $O(n\frac{kh}{w}B)$. (iii) TailApprox. Theorem 36 shows that the tail-approximation algorithm runs in time $O(n\frac{kh}{w}(\log n + \log\frac{x_{\max}}{x_{\min}}))$. Combining these three bounds gives the running time stated in the theorem. □

Note that for $B = O(k)$, the measurement bound gives $m = O(k\log\log\frac{k}{w})$, which is a significant improvement over the standard compressive sensing measurement bound $m = O(k\log\frac{n}{k})$. In fact, the bound for m is only a $\log\log\frac{k}{w}$ factor away from the information-theoretically optimal bound $m = O(k)$. We leave it as an open problem whether this spurious factor can be eliminated via a more refined analysis or algorithm.

9 Conclusions

We have introduced a new framework called *approximation-tolerant model-based compressive sensing*. Our framework consists of a range of algorithms for model-based compressive sensing that succeed even when the model-projection oracles are approximate. All our algorithms involve oracles that provide constant-factor approximations to both the "head" and "tail" versions of the model-projection problem. We have instantiated these algorithms for the Constrained Earth Mover Distance (CEMD) model. To achieve this, we have designed novel polynomial-time head- and tail-approximation oracles for the CEMD model based on graph optimization techniques. Leveraging these oracles and our framework results in nearly sample-optimal recovery schemes for signals belonging to this model.

Several avenues for future work remain. We have developed model-based recovery schemes that succeed with dense measurement matrices (AM-IHT, AM-CoSaMP), as well as sparse matrices (AM-IHT with RIP-1). An interesting question is whether model-based recovery can be extended to other classes of measurement matrices, such as subsampled Fourier matrices [45]. Also, the required sample-complexity m specified by Theorem 37 is a factor of $\log(\frac{B}{k}\log\frac{k}{w})$ away from the optimal $m = O(k)$, and it is possible that a different approach is needed to remove this log-factor. Finally, finding an efficient algorithm (or proving a computational hardness result) for *exact* projections into the CEMD model remains an open question.

References

[1] C. Hegde, P. Indyk, and L. Schmidt, "Approximation-tolerant model-based compressive sensing," in *Proc. ACM-SIAM Symp. Discrete Alg. (SODA)*, 2014.

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

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

[4] R. Kainkaryam, A. Bruex, A. Gilbert, J. Schiefelbein, and P. Woolf, "poolMC: Smart pooling of mRNA samples in microarray experiments," *BMC Bioinformatics*, vol. 11, no. 1, 2010.

[5] S. Muthukrishnan, "Data streams: Algorithms and applications," *Found. Trends Theor. Comput. Sci.*, vol. 1, no. 2, pp. 117–236, 2005.

[6] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 937–947, 2010.

[7] K. Do Ba, P. Indyk, E. Price, and D. Woodruff, "Lower Bounds for Sparse Recovery," in *Proc. ACM-SIAM Symp. Discrete Alg. (SODA)*, 2010.

[8] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich, "The Gelfand widths of ℓ_p-balls for $0 \leq p \leq 1$," *J. Complex.*, vol. 26, no. 6, pp. 629–640, 2010.

[9] B. Kasin, "Diameters of some finite-dimensional sets and classes of smooth functions.," *Math. USSR, Izv.*, vol. 11, pp. 317–333, 1977.

[10] A. Garnaev and E. Gluskin, "On widths of the Euclidean ball.," *Sov. Math., Dokl.*, vol. 30, pp. 200–204, 1984.

[11] E. Gluskin, "Norms of random matrices and widths of finite-dimensional sets.," *Math. USSR, Sb.*, vol. 48, pp. 173–182, 1984.

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

[13] Y. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[14] M. Duarte and Y. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. Sig. Proc.*, vol. 59, no. 9, pp. 4053–4085, 2011.

[15] N. Rao, B. Recht, and R. Nowak, "Universal measurement bounds for structured sparse signal recovery," in *Intl. Conf. Artificial Intel. Stat. (AISTATS)*, 2012, pp. 942–950.

[16] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, vol. 4, no. 1, pp. 1–106, 2012.

[17] M. Wainwright, "Structured regularizers for high-dimensional problems: Statistical and computational issues," *Annual Review of Statistics and Its Application*, vol. 1, no. 1, pp. 233–253, 2014.

[18] V. Cevher, P. Indyk, C. Hegde, and R. Baraniuk, "Recovery of clustered sparse signals from compressive measurements," in *Int. Conf. on Sampling Theory and Applications (SampTA)*, 2009.

[19] C. Hegde, M. Duarte, and V. Cevher, "Compressive sensing recovery of spike trains using a structured sparsity model," in *Sig. Proc. Adaptive Sparse Structured Rep. (SPARS)*, 2009.

[20] C. Carter and A. Thompson, "An exact tree projection algorithm for wavelets," *IEEE Signal Proc. Letters*, vol. 20, no. 11, pp. 1026–1029, 2013.

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

[22] T. Blumensath and M. Davies, "Iterative hard thresholding for compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 27, no. 3, pp. 265–274, 2009.

[23] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss, "Combining geometry and combinatorics: A unified approach to sparse signal recovery," in *Proc. Allerton Conf. on Comm., Contr., and Comp.*, 2008.

[24] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Springer Birkhäuser, 2013.

[25] B. Bah, L. Baldaserre, and V. Cevher, "Model-based sketching and recovery with expanders," in *Proc. ACM-SIAM Symp. Discrete Alg. (SODA)*, 2014.

[26] L. Schmidt, C. Hegde, and P. Indyk, "The Constrained Earth Mover Distance model, with applications to compressive sensing," in *Intl. Conf. on Sampling Theory and Appl. (SampTA)*, 2013.

[27] P. Indyk and I. Razenshteyn, "On model-based RIP-1 matrices," in *Intl. Coll. Automata, Lang. and Prog. (ICALP)*, 2013, (in particular, the updated version arXiv:1304.3604v3).

[28] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Inform. Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[29] A. Kyrillidis and V. Cevher, "Sublinear time, approximate model-based sparse recovery for all," *arXiv:1203.4746*, 2012.

[30] A. Kyrillidis and V. Cevher, "Combinatorial selection and least absolute shrinkage via the CLASH algorithm," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, Cambridge, MA, Jul. 2012.

[31] R. Giryes and M. Elad, "Iterative hard thresholding with near optimal projection for signal recovery," in *Intl. Conf. on Sampling Theory and Appl. (SampTA)*, 2013.

[32] M. Davenport, D. Needell, and M. Wakin, "Signal space CoSaMP for sparse recovery with redundant dictionaries," *IEEE Trans. Inform. Theory*, vol. 59, no. 10, pp. 6820–6829, 2013.

[33] R. Giryes and D. Needell, "Greedy signal space methods for incoherence and beyond," to appear in Appl. Comput. Harmon. Anal., 2014.

[34] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Sig. Proc.*, vol. 58, no. 9, pp. 4595–4607, 2010.

[35] M. Duarte, S. Sarvotham, D. Baron, M. Wakin, and R. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Asilomar Conf. Signals, Sys., Comput*, 2005.

[36] C. Hegde, P. Indyk, and L. Schmidt, "A fast approximation algorithm for tree-sparse recovery," in *Proc. IEEE Int. Symp. Inform. Theory (ISIT)*, 2014.

[37] C. Hegde, P. Indyk, and L. Schmidt, "Nearly linear-time model-based compressive sensing," in *Intl. Coll. Automata, Lang. and Prog. (ICALP)*, 2014.

[38] L. Schmidt, C. Hegde, P. Indyk, J. Kane, L. Lu, and D. Hohl, "Automatic fault localization using the Generalized Earth Movers Distance," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, 2014.

[39] T. Blumensath and M. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[40] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Const. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[41] K. Lee and Y. Bresler, "Admira: Atomic decomposition for minimum rank approximation," *IEEE Trans. Inform. Theory*, vol. 56, no. 9, pp. 4402–4416, 2010.

[42] E. Levina and P. Bickel, "The Earth Mover's distance is the Mallows distance: some insights from statistics," in *Proc. IEEE Intl. Conf. Comp. Vision (ICCV)*, 2001.

[43] R. Ahuja, T. Magnanti, and J. Orlin, *Network Flows: Theory, Algorithms, and Applications*, Prentice-Hall, Inc., 1993.

[44] T. Cormen, C. Stein, R. Rivest, and C. Leiserson, *Introduction to Algorithms*, McGraw-Hill Higher Education, 2nd edition, 2001.

[45] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Comm. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1171, 2008.