Approximation-Tolerant Model-Based Compressive Sensing

Chinmay Hegde, Piotr Indyk, Ludwig Schmidt

MIT

{chinmay,indyk,ludwigs}@mit.edu

Abstract

The goal of sparse recovery is to recover a k-sparse signal $x \in \mathbb{R}^n$ from (possibly noisy) linear measurements of the form $y = Ax$, where $A \in \mathbb{R}^{m \times n}$ describes the measurement process. Standard results in compressive sensing show that it is possible to recover the signal x from $m = O(k \log(n/k))$ measurements, and that this bound is tight. The framework of *model-based compressive sensing* [BCDH10] overcomes the lower bound and reduces the number of measurements further to $O(k)$ by limiting the supports of x to a subset $\mathcal{M}$ of the $\binom{n}{k}$ possible supports. This has led to many measurement-efficient algorithms for a wide variety of signal models, including block-sparsity and tree-sparsity.

Unfortunately, extending the framework to other, more general models has been stymied by the following obstacle: for the framework to apply, one needs an algorithm that, given a signal x, solves the following optimization problem *exactly* :

$$\arg\min_{\Omega \in \mathcal{M}} \|x_{[n]\setminus\Omega}\|_2$$

(here $x_{[n]\setminus\Omega}$ denotes the projection of x on coordinates not in Ω). However, an approximation algorithm for this optimization task is not sufficient. Since many problems of this form are not known to have exact polynomial-time algorithms, this requirement poses an obstacle for extending the framework to a richer class of models.

In this paper, we remove this obstacle and show how to extend the model-based compressive sensing framework so that it requires only *approximate* solutions to the aforementioned optimization problems. Interestingly, our extension requires the existence of approximation algorithms for both the maximization and the minimization variants of the optimization problem.

Further, we apply our framework to the *Constrained Earth Mover's Distance (CEMD)* model introduced in [SHI13], obtaining a sparse recovery scheme that uses significantly less than $O(k \log(n/k))$ measurements. This is the first non-trivial theoretical bound for this model, since the validation of the approach presented in [SHI13] was purely empirical. The result is obtained by designing a novel approximation algorithm for the maximization version of the problem and proving approximation guarantees for the minimization algorithm described in [SHI13].

1 Introduction

Over the last decade, a new "linear" approach for obtaining a succinct approximate representation of n-dimensional vectors (or signals) has been discovered. For any signal x, the representation is equal to Ax, where A is an $m \times n$ matrix, or possibly a random variable chosen from some distribution over such matrices. The vector Ax is often referred to as the *measurement vector* or *linear sketch* of x. Although m is typically much smaller than n, the sketch Ax often contains plenty of useful information about the signal x.

A particularly useful and well-studied problem is that of *robust sparse recovery*. We say that 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,[1] the recovery algorithm reports x^* such that:

$$\|x - x^*\|_2 \leq C\|e\|_2 \,. \tag{1}$$

Sparse recovery has a tremendous number of applications in areas such as compressive sensing of signals [CRT06, Don06], genetic data analysis [KBG+10], and data stream algorithms [Mut05, GI10].

It is known [GI10] that there exist matrices A and associated recovery algorithms that produce approximations x^* satisfying Equation (1) with a constant approximation factor C, and sketch length $m = O(k \log(n/k))$. It is also known that the bound on the number of measurements is asymptotically optimal for some constant C, see [BIPW10] and [FPRU10] (building on [GG84, Glu84, Kas77]). The necessity of the "extra" logarithmic factor multiplying k is quite unfortunate: the sketch length determines the "compression rate", and for large n any logarithmic factor can worsen that rate tenfold. However, more careful modeling offers a way to overcome the aforementioned limitation. In particular, after decades of research in signal modeling, signal processing researchers know that not all supports (i.e., sets of non-zero coordinates) are equally common. For example, if a signal is a function of time, large coefficients of the signal tend to occur consecutively. This phenomenon can be exploited by assuming that the support of the measured vector x belongs to a given "model" family of supports $\mathcal{M}_k$ (i.e., x is $\mathcal{M}_k$-*sparse*). The original k-sparse recovery problem corresponds to the case when $\mathcal{M}_k$ is the family of all k-subsets of $[n]$.

An elegant *model-based* sparse recovery scheme was recently provided in the work of Baraniuk et al. [BCDH10]. The scheme has the property that, for any "computationally tractable" family of supports of "small" size, it guarantees a near-optimal sketch length $m = O(k)$, i.e., without any logarithmic factors. The framework is general but relies on two model-specific conditions:

1. *Model-based Restricted-Isometry Property (RIP):* the mapping A must approximately preserve the norm of all $\mathcal{M}$-sparse vectors, and
2. *Model projection oracle:* the model must be supported by an efficient algorithm that, given an arbitrary vector x, finds the $\mathcal{M}$-sparse vector x' that is closest to x, i.e., minimizes the "tail" error $\|x - x'\|_2$.

By constructing mappings A satisfying (1) and algorithms satisfying (2), several algorithms were developed for a wide variety of signal models, including block-sparsity and tree-sparsity. In fact, it

[1] The robust sparse recovery problem is the most general version of sparse recovery. In particular, it subsumes the so-called *stable* sparse recovery problem, where the measured vector x is not assumed to be k-sparse, the measurement vector is equal to Ax, and the goal is to recover an approximation x^* such that $\|x-x^*\|_p \leq C(k) \min_{k\text{-sparse } x'} \|x-x'\|_q$ for some norm parameters p, q. This is because for the vector x' that minimizes the expression, we have $x = x'+(x-x')$ where x' is k-sparse, and the measurement vector is equal to $Ax = Ax' + A(x - x') = Ax' + e$. As a result we will not specifically consider stable sparse recovery in this paper, but refer the reader to Appendix B of [IP11] for details.

is known that the condition (1) holds for a much more general class of signal models [BCDH10]. Unfortunately, extending the whole framework to other models faces a considerable obstacle: for the framework to apply, the model projection algorithm has to be *exact*.[2] This excludes many useful design paradigms employed in *approximation algorithms*, such as greedy approaches, LP and SDP rounding, etc. As a result, most of the existing algorithms are based on exact dynamic programming [BCDH10], solving LPs without an integrality gap [HDC09], etc.

1.1 Our results

The contributions of this paper are two-fold. First, we extend the model-based compressive sensing framework so that it tolerates approximation algorithms. Second, we apply the new framework to a signal model called *Constrained Earth Mover's Distance Model*, obtaining the first theoretical bounds on its measurement complexity.

Approximate model projection oracles We extend the model-based framework to support *approximate* model projection oracles. Our extension allows approximation, but requires *two* oracles: the *tail* oracle, which approximates $\min_{\Omega \in \mathcal{M}} \|x_{[n]\setminus\Omega}\|_2$, as well as the *head* oracle, which approximates $\max_{\Omega \in \mathcal{M}} \|x_\Omega\|_2$. For any model, given these two oracles as well a matrix A that satisfies the model-based RIP, our framework leads to an algorithm satisfying Equation 1 (see Corollary 14 for details).

Sparse recovery for the Constrained Earth Mover's Distance Model We employ the new framework to obtain a model-based compressive sensing algorithm for the *Constrained Earth Mover's Distance (CEMD)* model introduced in [SHI13]. In this model, the signal coordinates form an $h \times w$ grid and the support of each column has size at most s, for $n = hw$ and $k = sw$. For each pair of consecutive columns (say c and c'), we define the EMD distance between them to be the minimum cost of matching the sets $\mathrm{supp}(c)$ and $\mathrm{supp}(c')$ viewed as point-sets on a line. The support set 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 3 for the formal definition.

We design approximation algorithms for both the head and the tail oracles for the CEMD model. Both algorithms have bicriterion approximation guarantees. Specifically, we develop:

1. An approximate tail oracle, which outputs a support set with the tail value at most $O(1)$ larger than the optimum and with the budget cost of $O(B)$ (Theorem 18), and
2. An approximate head oracle, which outputs a support set with the head value at least $\Omega(1)$ of the optimum and with the budget cost of $O(B \log k)$ (Theorem 16).

The tail oracle is obtained using min-cost max-flow techniques and Lagrangian relaxation. The head oracle is obtained using a greedy algorithm that iteratively selects s paths forming the matching with varying budgets. We then instantiate the approximate model-based framework with these algorithms to obtain a compressive sensing scheme for the CEMD model that uses $O(k \log(\frac{B}{k} \log(\frac{k}{w})))$ measurements. For slowly varying supports, i.e., $B = O(k)$, the bound specializes to $O(k \log \log(\frac{k}{w}))$.

1.2 Related work

There has been a large of body of work dedicated to algorithms for model-based compressive sensing (see, for example, the survey [DE11]). Unfortunately, the success of most of these algorithms relies on the availability of an exact model-projection oracle. The only works on approximate projection

[2]This fact might appear quite surprising, given that the framework ultimately produces an *approximation* algorithm. However, as we show in Appendix B, the framework provably leads to incorrect recovery if the model projection algorithm is not exact. Of course, this does not exclude the possibility that for *specific* models the original algorithm produces correct results, either in theory or in practice.

oracles that we are aware of either provide *additive* approximation guarantees [Blu11, KC12], or make very strong assumptions about the measurement matrix A or the projection oracle [GE13, DNW13]. Specifically, [Blu11] discusses a Projected Landweber-type method that succeeds even when the projection oracle is approximate; however, they assumed that the projection oracle provides an ϵ-*additive* tail approximation factor. Under such conditions, there exists an algorithm that returns a solution within an $O(\epsilon)$-neighborhood of the optimal solution. However, approximation algorithms with low additive approximation factors are rather rare. On the other hand, [KC12] assumes the existence of a variant of the approximate head oracle (called PMAP_ϵ), but provides approximation guarantees with an additive term of $O(\sqrt{\epsilon}\|x_\Omega\|)$ where Ω is the set of the k largest coefficients in x (cf. Theorem 4.3). The paper [GE13] present a sparse recovery algorithm that succeeds with multiplicative approximation guarantees. However, their framework uses only the tail oracle and therefore is subject to the lower bound outlined in Appendix B. In particular, their guarantees need to make very stringent assumptions on the singular values of the sensing matrix. Finally, while the paper [DNW13] also assumes the existence of multiplicative approximate oracles, our approach in comparison succeeds with considerably weaker assumptions.

The Constrained Earth Mover's Distance model was introduced in [SHI13]. The model was motivated by the task of reconstructing time sequences of spatially sparse signals,[3] e.g. seismic measurements. The paper introduced a tail oracle algorithm for the problem and empirically evaluated the performance of the scheme. Although the use of the oracle was heuristic, the experiments demonstrate substantial reduction in the number of measurements needed to recover slowly varying signals. In this paper we present *approximation guarantees* for the tail oracle from [SHI13], a novel algorithm for the head oracle, as well as the framework for showing that these two sub-routines yield a model-based compressive sensing scheme with a non-trivial measurement bound.

Another related paper is [IP11], whose authors propose the use of the EMD to measure the *approximation error* of the recovered signal in compressive sensing. In contrast, we are using the EMD to constrain the *support set* of the signals.

2 Preliminaries

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, $\text{supp}(x) \subseteq [n]$, is the set of indices with nonzero entries in x. For a matrix $X \in \mathbb{R}^{h \times w}$, the support $\text{supp}(X) \subseteq [h] \times [w]$ is the set of indices corresponding to nonzero entries. We denote the support of a column in X with $\text{col-supp}(X, c) = \{r \mid (r, c) \in X\}$.

Often, some prior information is available about the support of a sparse signal x; for example, in the case of "bursty" signals, the nonzeros of x may occur as a small number of blocks. A more general way to model such prior information is to consider all possible k-sparse signals with only a few permitted configurations of $\text{supp}(x)$. This restriction motivates the notion of a *structured sparsity model*, which is geometrically equivalent to a subset of the $\binom{n}{k}$ canonical subspaces of $\mathbb{R}^n$.

Definition 1 (Structured sparsity model[4])**.** *A structured sparsity model $\mathcal{M}_p \subseteq \mathbb{R}^n$ is the set of vectors $\mathcal{M}_p = \{x \in \mathbb{R}^n \mid \text{supp}(x) \subseteq S,\ S \in \mathbb{M}_p\}$, where $\mathbb{M}_p = \{\Omega_1, \ldots, \Omega_{a_p}\}$ is the set of allowed structured supports with $\Omega_i \subseteq [n]$. We call $a_p = |\mathbb{M}_p|$ the size of the model $\mathcal{M}_p$.*

The above definition utilizes the concept of a *model parameter*, $p \in \mathcal{P}$, where $\mathcal{P}$ is the set of possible parameters for a given model. This parameter is model dependent, and quantitatively

[3] There has been a substantial amount of work devoted to such signals (e.g., [VL10, DSB+05]). We refer the reader to [SHI13] for a more detailed discussion about the model and its applications.

[4] Definition 2 in [BCDH10].

encodes the additional structure of $\mathcal{M}_p$ (for example, for the case of block-sparsity, p can denote the block size). For two model parameters p and q, the structured sparsity model $\mathcal{M}_{p\oplus q}$ is defined by the set of supports $\mathbb{M}_{p\oplus q} = \{\Omega \cup \Gamma \,|\, \Omega \in \mathcal{M}_p \text{ and } \Gamma \in \mathcal{M}_q\}$.

The framework of model-based compressive sensing [BCDH10] leverages the above notion of a structured sparsity model to design robust sparse recovery schemes that improve upon existing approaches. Specifically, the framework states that it is possible to recover a structured-sparse signal $x \in \mathcal{M}_p$ from linear measurements $y = Ax + e$, provided that two conditions are satisfied: (i) the matrix A satisfies a type of 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 $\mathcal{M}_p$. Formally:

Definition 2 (Model-RIP[5])**.** *The matrix $A \in \mathbb{R}^{m\times n}$ has the (δ, p)-model-RIP if the following inequalities hold for all $x \in \mathcal{M}_p$:*

$$(1-\delta)\|x\|_2^2 \;\le\; \|Ax\|_2^2 \le (1+\delta)\|x\|_2^2 . \tag{2}$$

Definition 3 (Model projection oracle[6])**.** *A model projection oracle is a function $M : \mathbb{R}^n \times \mathcal{P} \to \mathbb{R}^n$ such that the following two properties hold for all $x \in R^n$ and $p \in \mathcal{P}$. (i)* Output model sparsity: *$M(x,p) \in \mathcal{M}_p$. (ii)* Optimal model projection: *$\|x - M(x,p)\|_2 = \min_{x'\in\mathcal{M}_p}\|x - x'\|_2$.*

The authors of [BCDH10] show that if the above two conditions are satisfied, then a simple modification of CoSaMP [NT09], or IHT [BD09a] (popular, iterative algorithms for sparse recovery) can be tailored to work for robust sparse recovery for *any arbitrary* structured sparsity model. We focus on IHT in this paper, and refer to the modified algorithm as *Model-IHT*. The potential benefit of such an approach stems on the model-RIP assumption: the following result indicates that with high probability, a large class of measurement matrices A can satisfy the model-RIP with a near-optimal number of rows:

Fact 4 ([BD09b, BCDH10])**.** *Let $\mathcal{M}_p$ be a structured sparsity model and let k be the size of the largest support in the model, i.e., $k = \max_{\Omega\in\mathbb{M}_p}|\Omega|$. Let $A \in \mathbb{R}^{m\times n}$ be a matrix with i.i.d. Gaussian entries. Then there is a constant c such that for fixed δ, any $t > 0$ and $m \ge c\,(k + \log a_p)$, A has the (δ, p)-model-RIP with probability at least $1 - e^{-t}$.*

If the number of permissible supports (equivalently, subspaces) a_p is asymptotically smaller than $\binom{n}{k}$, then m can be as small as $O(k)$ and this behavior is order-optimal. However, the efficiency of the above framework crucially depends on the running time of the model-projection oracle, since the overall recovery algorithm (Model-IHT) involves several invocations of the model projection oracle. See [DE11] for a discussion of several models that admit efficient projection algorithms.

3 The CEMD model

Before proceeding to our main results, we discuss a special structured sparsity model known as the *Constrained EMD* model [SHI13]. A key ingredient in the model is the Earth Mover's Distance (EMD), also known as the Wasserstein metric or Mallows distance [LB01]:

Definition 5 (EMD)**.** *The EMD of two sets $A, B \subset \mathbb{N}$ with $|A| = |B|$ is defined as* $\mathrm{EMD}(A,B) = \min_{\pi:A\to B}\sum_{a\in A}|a - \pi(a)|$*, where π ranges over all one-to-one mappings from A to B.*

Observe that $\mathrm{EMD}(A,B)$ is the cost of a min-cost matching between A and B. Now consider the case where the sets A and B correspond to the *supports* of two exactly k-sparse signals, so that

[5]Definition 3 in [BCDH10].
[6]Section 3.2 in [BCDH10].

$|A| = |B| = k$. In this case, the EMD not only measures how many indices change, but also how far the supported indices move. This intuition for pairs of signals can be generalized to an *ensemble* of sparse signals.

Definition 6 (Support-EMD of a matrix)**.** *Let $A \subseteq [h] \times [w]$ be the support of a matrix with exactly s-sparse columns, i.e., $|\text{col-supp}(A, c)| = s$ for $c \in [w]$. Then the EMD of A is defined as* $\text{EMD}(A) = \sum_{c=1}^{w-1} \text{EMD}(\text{col-supp}(A, c), \text{col-supp}(A, c+1))$.

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

The above definitions motivate a natural structured sparsity model that, in essence, 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 $n = h\,w$. 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 7 (Constrained EMD model)**.** *The Constrained EMD (CEMD) model is the structured sparsity model $\mathcal{M}_{k,B}$ defined by the set of supports $\mathbb{M}_{k,B} = \{\Omega \subseteq [h] \times [w] \,|\, \text{EMD}(\Omega) \leq B$ and $|\text{col-supp}(\Omega, c)| = s$ 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 4 in Appendix C illustrates the CEMD model with an example.

We show that the sum of two signals in the CEMD model also belongs to the CEMD model (with reasonably adjusted parameters). To see this, let $X, Y \in \mathbb{R}^{h\times w}$. Moreover, assume that $X \in \mathcal{M}_{k_1,B_1}$ and $Y \in \mathcal{M}_{k_2,B_2}$. Then $X + Y \in \mathcal{M}_{k_1+k_2,B_1+B_2}$. Each column of $X + Y$ is $\frac{k_1+k_2}{w}$ sparse. Also, we can use the matchings in X and Y to construct a matching for $X + Y$ with support-EMD at most $B_1 + B_2$. Further, this means that for our model, $\mathcal{M}_{p_1 \oplus p_2} \subseteq \mathcal{M}_{p_1+p_2}$.

Suppose that 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) $a_{k,B}$ in the model and invoke Fact 4. For simplicity, we will assume that $w = \Omega(\log h)$, i.e., the following bounds apply for all signals X, excluding very thin and very tall matrices X. The following result is novel:

Theorem 8. *The number of subspaces satisfies* $\log a_{k,B} = O\big(k \log \frac{B}{k}\big)$.

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 bins. (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 $\binom{h}{s}\binom{B+k-1}{k}2^k$, an upper bound on the number of

supports. Using the inequality $\binom{a}{b} \leq \left(\frac{a\,e}{b}\right)^b$, we get

$$\begin{aligned}\log a_{k,B} &\leq \log\left(\binom{h}{s}\binom{B+k-1}{k}2^k\right)\\ &\leq s\log\frac{h}{s} + k\log\frac{B+k}{k} + O(s+k)\\ &= O\left(k\log\frac{B}{k}\right).\end{aligned}$$

□

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 4, $m = O(\log a_{k,B} + k) = O(k)$. As mentioned above, 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 $O(k \log \frac{n}{k})$.

4 Approximate Model-IHT

As the second ingredient in our proposed sparse recovery scheme, we require a model projection oracle that, for any arbitrary signal x, returns a signal $x' \in \mathcal{M}_{k,B}$ with the optimal tail error. In Sections 5 and 6, we develop algorithms that perform such a projection; *however, they are only approximate and not necessarily optimal.* Therefore, we extend the model-based compressive sensing framework to work with approximate projection oracles (formalized 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. Since this extension can be of independent interest, we present the results in a very general setting.

Definition 9 (Head approximation oracle)**.** *Let $c \in \mathbb{R}$ and $f : \mathcal{P} \to \mathcal{P}$. A (c, f)-head approximation oracle is a function $H : \mathbb{R}^n \times \mathcal{P} \to \mathbb{R}^n$ such that the following two properties hold for all $x \in \mathbb{R}^n$ and $p \in \mathcal{P}$:*
Output model sparsity: $H(x, p) = x_\Omega$ for some $\Omega \in \mathbb{M}_{f(p)}$.
Head approximation: $\|H(x, p)\|_2 \geq c\|x_\Omega\|_2$ for all $\Omega \in \mathbb{M}_p$.

Definition 10 (Tail approximation oracle)**.** *Let $c \in \mathbb{R}$ and $f : \mathcal{P} \to \mathcal{P}$. A (c, f)-tail approximation oracle is a function $T : \mathbb{R}^n \times \mathcal{P} \to \mathbb{R}^n$ such that the following two properties hold for any $x \in \mathbb{R}^n$ and $p \in \mathcal{P}$:*
Output model sparsity: $T(x, p) = x_\Omega$ for some $\Omega \in \mathbb{M}_{f(p)}$.
Tail approximation: $\|x - T(x, k)\|_2 \leq c\|x - x'\|_2$ for all $x' \in \mathcal{M}_p$.

We sometimes write $H_p(x)$ instead of $H(x, p)$, and $T_p(x)$ instead of $T(x, p)$ for clarity of presentation. 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 $\mathcal{M}_p$, if $x \in \mathcal{M}_p$ then $T(x, p) = x$ regardless of the choice of tail approximation oracle; however, $H(x, p)$ need not necessarily return the signal x. An important feature of the above definitions of approximate oracles is that they allow for projections into *larger* models. In particular, the oracle can potentially return a signal that belongs to a model $\mathbb{M}_{f(p)}$ specified by the model parameter $f(p)$; for instance, a tail-approximation oracle for the CEMD model with parameters (k, B) is allowed to return a signal with parameters $(2k, 2B)$. We exploit this feature in our algorithms below.

Equipped with these notions of approximate oracles, we introduce a sparse recovery algorithm for model-based compressive sensing that we call *Approximate Model-IHT* (AM-IHT); see Algorithm 1.

Algorithm 1 Approximate model-IHT

function AM-IHT(y,A,p,t)
  $x^1 \leftarrow 0$
  for $i \leftarrow 1, \ldots, t$ **do**
    $x^{i+1} \leftarrow T_p(x^i + H_{p \oplus f_T(p)}(A^T(y - Ax^i)))$
  return x^{t+1}

Note that we use both a (c_H, f_H)-head approximation oracle H and a (c_T, f_T)-tail approximation oracle T in every iteration of AM-IHT. This is in contrast with the usual version of IHT (and its model-based extension) which uses only one oracle projection.[7] Our central result (Theorem 13) states that if a matrix A satisfies the model-RIP with parameters $(\delta, f_H(p \oplus f_T(p)) \oplus p \oplus f_T(p))$, and approximation oracles H and T are available, 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 $x \in \mathcal{M}_p$. (ii) $y = Ax + e$ for an arbitrary $e \in \mathbb{R}^m$ (the measurement noise). (iii) A has (δ, t)-model-RIP for $t = f_H(p \oplus f_T(p)) \oplus p \oplus f_T(p)$. Moreover, we define the following quantities as shorthands: (i) $r^i = x - x_i$. (ii) $a^i = x^i + H_{p \oplus f_T(p)}(A^T(y - Ax^i))$. (iii) $b^i = A^T(y - Ax^i)$. (iv) $\Omega = \operatorname{supp}(r^i)$., and (v) $\Gamma = \operatorname{supp}(H_{p \oplus f_T(p)}(b^i))$.

As a preliminary lemma, we show that we can use the RIP of A on relevant vectors.

Lemma 11. *For all $x \in \mathbb{R}^n$ with $\operatorname{supp}(x) \in \Omega \cup \Gamma$ we have*

$$(1 - \delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta)\|x\|_2^2 .$$

Proof. By the definition of T, we have $\operatorname{supp}(x^i) \in \mathbb{M}_{f_T(p)}$. Since $\operatorname{supp}(x) \in \mathbb{M}_p$, we have $\operatorname{supp}(x - x^i) \in \mathbb{M}_{p \oplus f_T(p)}$ and hence $\Gamma \in \mathbb{M}_{p \oplus f_T(p)}$. Moreover, $\operatorname{supp}(H_{p \oplus f_T(p)}(b^i)) \in \mathbb{M}_{f_H(p \oplus f_T(p))}$ by the definition of H. Therefore $\Omega \cup \Gamma \in \mathbb{M}_{f_H(p \oplus f_T(p)) \oplus p \oplus f_T(p)}$, which allows us to use the model-RIP of A on x with $\operatorname{supp}(x) \in \Omega \cup \Gamma$. □

As a result of Lemma 11, we can use the standard consequences of the RIP, such as the approximate orthogonality defined in Section A.1. Just as in IHT, we use the residual proxy $A^T(y - Ax^i)$ as update in each iteration. We show that H preserves the relevant part of the residual proxy.

Lemma 12.

$$\left\| A^T_{\Omega \setminus \Gamma} A r^i \right\|_2 \leq \frac{\sqrt{1 - c_H^2}\,(1 + \delta) + \delta}{c_H} \left\| r^i \right\|_2 + \sqrt{1 + \delta} \left(\frac{\sqrt{1 - c_H^2} + 1}{c_H} + 1 \right) \|e\|_2 .$$

Proof. For the rest of this proof, we will denote $\|\cdot\|$ to denote the ℓ_2-norm. The head approximation

[7] In fact, we show that signal recovery using conventional model-IHT *fails* in cases where, instead of an exact projection oracle, only a tail approximation oracle is employed in every step; see Appendix B.

guarantee of H gives us the following series of results:

$$\begin{aligned}
\left\|b^i_\Gamma\right\|^2 &\geq c_H^2\left\|b^i_\Omega\right\|^2 \\
\left\|b^i_{\Gamma\cap\Omega}\right\|^2 + \left\|b^i_{\Gamma\setminus\Omega}\right\|^2 &\geq c_H^2\left\|b^i_{\Gamma\cap\Omega}\right\|^2 + c_H^2\left\|b^i_{\Omega\setminus\Gamma}\right\|^2 \\
\frac{(1-c_H^2)\left\|b^i_{\Gamma\cap\Omega}\right\|^2 + \left\|b^i_{\Gamma\setminus\Omega}\right\|^2}{c_H^2} &\geq \left\|b^i_{\Omega\setminus\Gamma}\right\|^2 \\
\frac{\|\sqrt{1-c_H^2}\,b^i_{\Gamma\cap\Omega} + b^i_{\Gamma\setminus\Omega}\|}{c_H} &\geq \left\|b^i_{\Omega\setminus\Gamma}\right\|.
\end{aligned}$$

We then expand $b^i = A^T(y - Ax_i)$ and apply consequences of the RIP, as well as the triangle inequality, several times:

$$\begin{aligned}
&\frac{1}{c_H}\|\sqrt{1-c_H^2}\,A^T_{\Gamma\cap\Omega}Ar^i + \sqrt{1-c_H^2}\,A^T_{\Gamma\cap\Omega}e + \\
&A^T_{\Gamma\setminus\Omega}Ar^i + A^T_{\Gamma\setminus\Omega}e\| \geq \left\|A^T_{\Omega\setminus\Gamma}Ar^i + A^T_{\Omega\setminus\Gamma}e\right\|, \\
&\frac{\sqrt{1-c_H^2}}{c_H}\|A^T_{\Gamma\cap\Omega}Ar^i\| + \frac{\sqrt{1-c_H^2}}{c_H}\|A^T_{\Gamma\cap\Omega}e\| + \\
&\frac{1}{c_H}\|^T_{\Gamma\setminus\Omega}Ar^i\| + \frac{1}{c_H}\left\|A^T_{\Gamma\setminus\Omega}e\right\| \geq \left\|A^T_{\Omega\setminus\Gamma}Ar^i\right\| - \left\|A^T_{\Omega\setminus\Gamma}e\right\|, \\
&\frac{\sqrt{1-c_H^2}\,(1+\delta)}{c_H}\|r^i\| + \frac{\sqrt{1-c_H^2}\,\sqrt{1+\delta}}{c_H}\|e\| + \\
&\frac{\delta}{c_H}\|r^i\| + \frac{\sqrt{1+\delta}}{c_H}\|e\| \geq \left\|A^T_{\Omega\setminus\Gamma}Ar^i\right\| - \sqrt{1+\delta}\|e\|.
\end{aligned}$$

Rearranging and grouping terms gives the statement of the lemma:

$$\begin{aligned}
\left\|A^T_{\Omega\setminus\Gamma}Ar^i\right\| \leq\ & \frac{\sqrt{1-c_H^2}\,(1+\delta)+\delta}{c_H}\|r^i\| \\
&+ \sqrt{1+\delta}\left(\frac{\sqrt{1-c_H^2}+1}{c_H}+1\right)\|e\|.
\end{aligned}$$

□

We now prove the main theorem: geometric convergence for approximate model-IHT.

Theorem 13 (Convergence of AM-IHT)**.**

$$\begin{aligned}
\left\|r^{i+1}\right\|_2 \leq\ & (1+c_T)\left(\frac{\sqrt{1-c_H^2}\,(1+\delta)+\delta}{c_H}+2\delta\right)\left\|r^i\right\|_2 \\
&+ (1+c_T)\sqrt{1+\delta}\left(\frac{\sqrt{1-c_H^2}+1}{c_H}+4\right)\|e\|_2.
\end{aligned}$$

Proof. Again, we use $\|\cdot\|$ to denote the ℓ_2-norm. The triangle inequality gives us

$$\begin{aligned}\left\|r^{i+1}\right\| &= \left\|x - x^{i+1}\right\| \\ &\leq \left\|x - a^i\right\| + \left\|x^{i+1} - a^i\right\| \\ &\leq (1 + c_T)\left\|x - a^i\right\|,\end{aligned}$$

where the last line follows because T is a (c_T, f_T)-tail approximation oracle. We now bound $\left\|x - a^i\right\|$:

$$\begin{aligned}\left\|x - a^i\right\| &= \left\|x - x^i - H_{p \oplus f_T(p)}(b^i)\right\| \\ &= \left\|r^i - H_{p \oplus f_T(p)}(b^i)\right\| \\ &\leq \left\|r^i - b^i_\Omega\right\| + \left\|H_{p \oplus f_T(p)}(b^i) - b^i_\Omega\right\|.\end{aligned}$$

Looking at each term individually, we have:

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

and

$$\begin{aligned}&\left\|H_{p \oplus f_T(p)}(b^i) - b^i_\Omega\right\| = \left\|b^i_\Gamma - b^i_\Omega\right\| \\ &= \left\|A^T_\Gamma A r^i - A^T_\Gamma e - A^T_\Omega A r^i - A^T_\Omega e\right\| \\ &\leq \left\|A^T_\Gamma A r^i - A^T_\Omega A r^i\right\| + 2\sqrt{1+\delta}\|e\| \\ &= \left\|A^T_{\Gamma\setminus\Omega} A r^i - A^T_{\Omega\setminus\Gamma} A r^i\right\| + 2\sqrt{1+\delta}\|e\| \\ &\leq \left\|A^T_{\Gamma\setminus\Omega} A r^i\right\| + \left\|A^T_{\Omega\setminus\Gamma} A r^i\right\| + 2\sqrt{1+\delta}\|e\| \\ &\leq \delta\left\|r^i\right\| + \left\|A^T_{\Omega\setminus\Gamma} A r^i\right\| + 2\sqrt{1+\delta}\|e\|.\end{aligned}$$

Using Lemma 12 gives us:

$$\begin{aligned}&\|H_{p \oplus f_T(p)}(b^i) - b^i_\Omega\| \\ &\leq \left(\frac{\sqrt{1 - c_H^2}\,(1+\delta) + \delta}{c_H} + \delta\right)\left\|r^i\right\| \\ &+ \sqrt{1+\delta}\left(\frac{\sqrt{1 - c_H^2} + 1}{c_H} + 3\right)\|e\|.\end{aligned}$$

Combining the inequalities above, we get

$$\begin{aligned}\left\|r^{i+1}\right\| &\leq (1 + c_T)\left(\frac{\sqrt{1 - c_H^2}\,(1+\delta) + \delta}{c_H} + 2\delta\right)\left\|r^i\right\| \\ &+ (1 + c_T)\sqrt{1+\delta}\left(\frac{\sqrt{1 - c_H^2} + 1}{c_H} + 4\right)\|e\|.\end{aligned}$$

□

Corollary 14. *For $\delta = 0.01$, $c_T = 1.5$ and $c_H = 0.95$ we get*

$$\|x - \text{AM-IHT}(y, A, p, t)\|_2 \leq 0.91^t \|x\|_2 + 150.34\|e\|_2 .$$

Proof. We iterate Theorem 13 and use $13.53 \sum_{i=0}^{\infty} 0.91^i \leq 150.34$. □

These results show that AM-IHT exhibits an overall recovery guarantee that is comparable to the existing model-based compressive sensing results of [BCDH10], despite using only approximate projection oracles. This has the potential to significantly extend the scope of efficient model-based sparse recovery methods. We instantiate this in the context of the CEMD model as follows.

5 Head Approximation Algorithm

First, we develop a head approximation oracle 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 $\|H_{k,B}(x)\|_2 = \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_2$. However, this appears to be a hard problem. Instead, we propose an efficient, greedy algorithm satisfying the (somewhat looser) properties of a head approximation oracle (Definition 9). Specifically, we develop an algorithm that performs the following task: given an arbitrary signal x and an approximation ratio c, find a support $\Omega \in \mathbb{M}_{O(k),O(B \log k)}$ such that $\|x_\Omega\|_2 \geq c \max_{\Gamma \in \mathbb{M}_{k,B}} \|x_\Gamma\|_2$.

As before, we will interpret our signal x as a matrix $X \in \mathbb{R}^{h \times w}$. For a signal $x \in \mathcal{M}_{k,B}$, we may interpret each support of x as a set of $s = k/w$ paths from the leftmost to the rightmost column in X. Hence the goal of our algorithm is to find a set of s such paths that cover a large amount of amplitudes in the signal. Let OPT denote the largest amplitude sum achievable with a support in $\mathbb{M}_{k,B}$, i.e., $OPT = \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_1$. Our method proceeds as follows. We first describe a scheme that allows us to get a constant fraction of the optimal amplitude sum OPT. Then, we repeat this algorithm several times in order to boost the approximation ratio while increasing the sparsity and EMD budget of the result only moderately.

Consider the closely related problem where $X_{i,j} \geq 0$ for all i, j and we are interested in the ℓ_1-guarantee

$$\|x_\Omega\|_1 \geq c \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_1 . \tag{3}$$

We can easily convert the input signal x into a matrix satisfying these constraints by squaring each amplitude. This modification allows us to simply add coefficient amplitudes along paths in the analysis of the head approximation algorithm.

Definition 15 (Path in a matrix)**.** *Given a matrix $X \in \mathbb{R}^{h \times w}$, a path $p \subseteq [h] \times [w]$ is a set of w locations in X with one location per column, i.e., $|p| = w$ and $\bigcup_{(i,j) \in p} j = [w]$. The weight of p is the sum of amplitudes on p, i.e., $w_X(p) = \sum_{(i,j) \in p} X_{i,j}$. The EMD of p is the sum of the EMDs between locations in neighboring columns. Let $j_1, \ldots, j_w$ be the locations of p in columns 1 to w. Then,* $\text{EMD}(p) = \sum_{i=1}^{w-1} |j_i - j_{i+1}|$.

Trivially, we have that a path p in X is a support with $w_X(p) = \|X_p\|_1$ and $\text{EMD}(p) = \text{EMD}(\text{supp}(X_p))$. Therefore, we can iteratively build a support Ω by finding s paths in X. Algorithm 2 contains the description of HeadApproxBasic. We show that HeadApproxBasic finds a constant fraction of the amplitude sum of the best support while only moderately increasing the size of the model. For simplicity, denote $w(p) := w_X(p)$, and $w^{(i)}(p) := w_{X^{(i)}}(p)$. We obtain the result:

Algorithm 2 Basic head approximation algorithm

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

Theorem 16. *Let Ω be the support returned by* HEADAPPROXBASIC. *Let $B' = \lceil H_s \rceil B$, where H_s is the s-th harmonic number. Then $\Omega \in \mathcal{M}_{k,B'}$ and $\|X_\Omega\|_1 \geq \frac{1}{4} OPT$. Moreover,* HEADAPPROXBASIC *runs in $O(s\,n\,B\,h)$ time.*

Proof. We can always decompose Ω_{OPT} into s disjoint paths in A. Let $p_1, \ldots, p_s$ be such a decomposition with $\mathrm{EMD}(p_1) \geq \mathrm{EMD}(p_2) \geq \ldots \geq \mathrm{EMD}(p_s)$. Note that $\mathrm{EMD}(p_i) \leq \lfloor \frac{B}{i} \rfloor$: otherwise $\sum_{j=1}^{i} \mathrm{EMD}(p_i) > B$ and since $\mathrm{EMD}(\Omega_{OPT}) \leq B$ this would be a contradiction. Since Ω is the union of s paths in A, Ω has column-sparsity s. Moreover, we have $\mathrm{EMD}(\Omega) = \sum_{i=1}^{s} \mathrm{EMD}(q_i) \leq \sum_{i=1}^{s} \lfloor \frac{B}{i} \rfloor \leq \lceil H_s \rceil B$. Therefore, $\Omega \in \mathcal{A}_{k,B'}$.

When finding path q_i in $X^{(i)}$, there are two cases: (i) either $w^{(i)}(p_i) \leq \frac{1}{2} w(p_i)$, i.e., the paths $q_1, \ldots, q_{i-1}$ have already covered more than half of the amplitude sum of p_i in X; (ii) or, $w^{(i)}(p_i) > \frac{1}{2} w(p_i)$, i.e., there is still more than half of the amplitude sum of p_i remaining in $X^{(i)}$. Since $\mathrm{EMD}(p_i) \leq \lfloor \frac{B}{i} \rfloor$, the path p_i is a candidate when searching for the optimal q_i and hence we find a path q_i with $w^{(i)}(q_i) > \frac{1}{2} w(p_i)$. Let $C = \{i \in [s] \mid \text{case (i) holds for } q_i\}$ and $D = \{i \in [s] \mid \text{case (ii) holds for } q_i\}$ (note that $C = [s] \setminus D$). Then we have

$$\begin{aligned} \|A_\Omega\|_1 &= \sum_{i=1}^{s} w^{(i)}(q_i) = \sum_{i \in C} w^{(i)}(q_i) + \sum_{i \in D} w^{(i)}(q_i) \\ &\geq \sum_{i \in D} w^{(i)}(q_i) \geq \frac{1}{2} \sum_{i \in D} w(p_i). \end{aligned} \tag{4}$$

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

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

and

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

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

$$\|X_\Omega\|_1 = \sum_{i=1}^{s} w^{(i)}(q_i) \geq \sum_{i \in C} \sum_{(u,v) \in E_i} X_{u,v} \geq \frac{1}{2} \sum_{i \in C} w(p_i)\,. \tag{5}$$

Combining Equations 4 and 5 gives:

$$2\|X_\Omega\|_1 \geq \frac{1}{2}\sum_{i\in C} w(p_i) + \frac{1}{2}\sum_{i\in D} w(p_i) = \frac{1}{2}OPT, \text{ i.e.,}$$
$$\|X_\Omega\|_1 \geq \frac{1}{4}OPT.$$

Further, observe that the running time of HeadApproxBasic depends on the running time of finding a path in a matrix X with maximum weight for a given EMD budget. Such a path search can be performed by *dynamic programming* over a graph with $whB = nB$ states. We have one state for each combination of location in X and amount of EMD budget currently used. At each state, we store the largest weight achieved by a path ending at the corresponding location in X and using the corresponding amount of EMD budget. 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 $O(nBh)$. Since we repeat this procedure s times, the overall time complexity of HeadApproxBasic is $O(snBh)$. □

Finally, we can use HeadApproxBasic to get a head approximation guarantee $\|x_\Omega\|_1 \geq c \max_{\Omega\in\mathbb{M}_{k,B}} \|x_\Omega\|_1$ for arbitrary $c < 1$. We achieve this by running HeadApproxBasic several times to get a progressively larger support that contains a larger fraction of OPT. We call the resulting algorithm HeadApprox. See Theorem 23 in Appendix A.2 for a rigorous proof.

6 Tail Approximation Algorithm

Next, we develop a tail approximation oracle for the CEMD model. Similar to before, we study the ℓ_1-version of the tail approximation problem: given an arbitrary signal x and an approximation ratio c, find a support $\Omega \in \mathbb{M}_{k,O(B)}$ such that

$$\|x - x_\Omega\|_1 \leq c \min_{\Gamma\in\mathbb{M}_{k,B}} \|x - x_\Gamma\|_1 . \tag{6}$$

Note that we allow a constant factor increase in the EMD budget of the result. The algorithm that we develop is precisely the graph-based approach initially proposed in [SHI13]; however, our analysis here is rigorous and novel. The core element of the algorithm is the notion of a *flow network*.

Definition 17 (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 *are 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 for $i \in [h]$.*
- *The* capacity *on every edge and node is 1.*
- *The* cost *of each node $v_{i,j}$ is $-|X_{i,j}|$. 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 and the demand at the sink is s.*

Figure 1 in Appendix C illustrates this definition with an example. Intuitively, a set of disjoint paths through the network $G_{X,k,\lambda}$ corresponds to supports in X. Therefore, for any fixed value of λ, a standard min-cost max-flow optimization through the flow network reveals a subset S of the nodes that (i) corresponds to a support with exactly s indices per column, and (ii) minimizes

$-\|X_\Omega\|_1 + \lambda \text{EMD}(\Omega)$ for different choices of supports Ω. In other words, the min-cost flow solves a *Lagrangian relaxation* of the original problem (6).

A crucial problem 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. Each support Ω maps to a single point in a 2D plane defined by EMD-cost in the x-direction and size of the tail in the y-direction, while each choice of λ defines a line with slope $-\lambda$. The union of all possible choices of λ defines a *convex hull* in the plane. A geometric perspective of this problem is illustrated in Figure 5 (Appendix C).

Therefore, finding min-cost flows corresponding to different choices of λ enables us to explore the convex hull of achievable supports. See Algorithm 4 (TailApprox) in Appendix A.3, which efficiently performs this exploration via a binary search over λ. While the best support for a particular target support-EMD B does not necessarily lie on the convex hull, we can show that we still get a near-optimal support with support-EMD $O(B)$. In particular, we have the following results with proofs in Appendix A.3.

Theorem 18. *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 EMD at most B, i.e., $OPT = \min_{\Gamma \in \mathbb{M}_{k,B}} \|X - X_\Gamma\|_1$. Then at least one of the following two guarantees holds for Ω: (i) either $B \leq \text{EMD}(\Omega) \leq dB$ and $\|X - X_\Omega\|_1 \leq OPT$, or (ii) $\text{EMD}(\Omega) \leq B$ and $\|X - X_\Omega\|_1 \leq (1+\delta)\frac{d}{d-1} OPT$. Moreover,* TailApprox *runs in $O(snh \log \frac{\|X\|_1 n}{x_{\min}\delta})$ time, where $x_{\min} = \min_{|X_{i,j}|>0} |X_{i,j}|$.*

Corollary 19. *Let $\delta > 0$ and $d = 1 + \frac{1}{c^2/(1+\delta)-1}$. Then* TailApprox *is a $(c, (k, dB))$-tail approximation algorithm.*

To summarize, the algorithm proposed in [SHI13] satisfies the criteria of a tail approximation oracle. This, in conjunction with the head approximation oracle proposed in Section 5, gives a fully developed sparse recovery scheme for the CEMD model, as described below.

7 Compressive Sensing with the CEMD Model

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 method that significantly reduces the number of measurements necessary for recovering signals in the CEMD model. The main result is the following theoretical guarantee, with the proof relegated to Appendix A.4.

Theorem 20. *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 $O(snh \log \frac{\|x\|_2}{\|e\|_2}(B + d\log(\|x\|_2 n)))$ if x, A and e are specified with at most d bits of precision.*

Observe that for $B = O(k)$, 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).

Acknowledgements

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.

References

[BCDH10] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Trans. Inform. Theory*, 56(4):1982–2001, 2010.

[BD09a] T. Blumensath and M. Davies. Iterative hard thresholding for compressed sensing. *Applied Comput. Harmonic Anal. (ACHA)*, 27(3):265–274, 2009.

[BD09b] T. Blumensath and M. Davies. Sampling theorems for signals from the union of finite-dimensional linear subspaces. *IEEE Trans. Inform. Theory*, 55(4):1872–1882, 2009.

[BIPW10] K. Do Ba, P. Indyk, E. Price, and D. Woodruff. Lower Bounds for Sparse Recovery. In *Proc. Symp. Discrete Algorithms (SODA)*, 2010.

[Blu11] T. Blumensath. Sampling and reconstructing signals from a union of linear subspaces. *IEEE Trans. Inform. Theory*, 57(7):4660–4671, 2011.

[CRT06] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.

[DE11] M. Duarte and Y. Eldar. Structured compressed sensing: From theory to applications. *IEEE Trans. Sig. Proc.*, 59(9):4053–4085, 2011.

[DNW13] M. Davenport, D. Needell, and M. Wakin. Signal space CoSaMP for sparse recovery with redundant dictionaries. To appear in IEEE Trans. Inform. Theory, 2013.

[Don06] D. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[DSB+05] 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.

[FPRU10] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of ℓ_p-balls for $0 \leq p \leq 1$. *J. Complex.*, 26(6):629–640, 2010.

[GE13] R. Giryes and M. Elad. Iterative hard thresholding with near optimal projection for signal recovery. In *10th Intl. Conf. on Sampling Theory and Appl. (SAMPTA)*, 2013.

[GG84] A. Garnaev and E. Gluskin. On widths of the Euclidean ball. *Sov. Math., Dokl.*, 30:200–204, 1984.

[GI10] A. Gilbert and P. Indyk. Sparse recovery using sparse matrices. *Proceedings of the IEEE*, 98(6):937–947, 2010.

[Glu84] E. Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. USSR, Sb.*, 48:173–182, 1984.

[HDC09] 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.

[IP11] P. Indyk and E. Price. K-median clustering, model-based compressive sensing, and sparse recovery for Earth Mover Distance. In *Proc. ACM Symp. Theory of Comput.*, 2011.

[Kas77] B. Kasin. Diameters of some finite-dimensional sets and classes of smooth functions. *Math. USSR, Izv.*, 11:317–333, 1977.

[KBG+10] R. Kainkaryam, A. Bruex, A. Gilbert, J. Schiefelbein, and P. Woolf. poolMC: Smart pooling of mRNA samples in microarray experiments. *BMC Bioinformatics*, 11(1), 2010.

[KC12] A. Kyrillidis and V. Cevher. Sublinear time, approximate model-based sparse recovery for all. *arXiv:1203.4746*, 2012.

[LB01] 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)*, volume 2, pages 251–256, 2001.

[Mut05] S. Muthukrishnan. Data streams: Algorithms and applications. *Found. Trends Theor. Comput. Sci.*, 1(2):117–236, 2005.

[NT09] D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied Comput. Harmonic Anal. (ACHA)*, 26(3):301–321, 2009.

[SHI13] L. Schmidt, C. Hegde, and P. Indyk. The Constrained Earth Mover Distance model, with applications to compressive sensing. In *10th Intl. Conf. on Sampling Theory and Appl. (SAMPTA)*, 2013.

[VL10] N. Vaswani and W. Lu. Modified-CS: Modifying compressive sensing for problems with partially known support. *IEEE Trans. Sig. Proc.*, 58(9):4595–4607, 2010.

Algorithm 3 Head approximation algorithm

function HEADAPPROX(X, k, B, c)
  $d \leftarrow \left\lceil \frac{4c}{1-c} \right\rceil$
  $X^{(1)} \leftarrow A$
  for $i \leftarrow 1, \ldots, d$ **do**
    $\Omega_i \leftarrow$ HEADAPPROXBASIC($X^{(i)}, k, B$)
    $X^{(i+1)} \leftarrow X^{(i)}$
    $X^{(i+1)}_{\Omega_i} \leftarrow 0$
  return $\bigcup_{i=1}^{d} \Omega_i$

Appendices

A Proofs

A.1 Near-isometry properties

A matrix with the RIP behaves like a near-isometry when restricted to a small set of columns and / or rows. The following properties are a direct result of the RIP.

Fact 21 (from Section 3 in [NT09])**.** *Let $A \in \mathbb{R}^{m \times n}$ be a matrix with (δ, p)-model-RIP. Let Ω be a support in the model, i.e., $\Omega \in \mathbb{M}_p$. Then the following properties hold for all $x \in \mathbb{R}^n$.*

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

Due to this near-isometry, a matrix with RIP is also "almost orthogonal" when restricted to a small set of columns and / or rows. The following property is therefore known as *approximate orthogonality.*

Fact 22 (from Section 3 in [NT09])**.** *Let $A \in \mathbb{R}^{m \times n}$ be a matrix with (δ, p)-model-RIP. Let Ω and Γ be two disjoint supports with their union in the model, i.e., $\Omega \cup \Gamma \in \mathbb{M}_p$. Then the following property holds for all $x \in \mathbb{R}^n$:*

$$\left\|A_\Omega^T A_\Gamma x\right\|_2 \leq \delta\|x\|_2 . \tag{7}$$

A.2 Head approximations

In this section, we use HEADAPPROXBASIC to get a head approximation guarantee

$$\|x_\Omega\|_1 \geq c \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|_1 \tag{8}$$

for arbitrary $c < 1$. We achieve this by running HEADAPPROXBASIC several times to get a larger support that contains a larger fraction of OPT. We call the resulting algorithm HEADAPPROX (see Algorithm 3). We use $d = \left\lceil \frac{4c}{1-c} \right\rceil$ as a shorthand throughout the analysis.

Theorem 23. *Let Ω be the support returned by* HEADAPPROXBASIC*. Let $k' = dk$ and $B' = d\lceil H_s \rceil B$. Then $\Omega \in \mathcal{M}_{k',B'}$ and $\|A_\Omega\|_1 \geq cOPT$.*

Proof. From Theorem 16 we know that for each i, $\Omega_i \in \mathcal{M}_{k,\lceil H_s \rceil B}$. Since $\Omega = \bigcup_{i=1}^{d} \Omega_i$, we have $\Omega \in \mathcal{M}_{k',B'}$. Before each call to HeadApproxBasic, at least one of the following two cases holds:

Case 1: $\|X^{(i)}_{\Omega_{OPT}}\|_1 < (1-c)OPT$. So the supports Ω_j found in previous iterations already cover amplitudes with a sum of at least $cOPT$.

Case 2: $\|X^{(i)}_{\Omega_{OPT}}\|_1 \geq (1-c)OPT$. Since Ω_{OPT} is a candidate solution with parameters k and B in $X^{(i)}$, we have that $\|X^{(i)}_{\Omega_i}\|_1 \geq \frac{1-c}{4}OPT$.

Consequently, after d iterations of the for-loop in HeadApprox, one of the following two cases holds:

Case A: Case 1 holds for at least one iteration. Hence $\|X_\Omega\|_1 \geq cOPT$.

Case B: Case 2 holds in all d iterations. Since we set $X^{(i+1)}_{\Omega_i} \leftarrow 0$ in each iteration, we have $\left\|X^{(i)}_{\Omega_j}\right\| = 0$ for all $j < i$. In particular, this means that $\left\|X^{(i)}_{\Omega_i}\right\|_1 + \left\|X^{(i+1)}_{\Omega_{i+1}}\right\|_1 = \left\|X^{(i)}_{\Omega_i \cup \Omega_{i+1}}\right\|_1$ and hence

$$\begin{aligned}\|X_\Omega\|_1 &= \sum_{i=1}^{d} \left\|X^{(i)}_{\Omega_i}\right\|_1 \\ &\geq \left\lceil \frac{4c}{1-c} \right\rceil \frac{1-c}{4} OPT \\ &\geq cOPT.\end{aligned}$$

So in both cases A and B, at the end of the algorithm we have $\|X_\Omega\|_1 \geq cOPT$. □

Corollary 24. HeadApprox *is a* $\left(c, \left(\left\lceil \frac{4c^2}{1-c^2} \right\rceil k, \left\lceil \frac{4c^2}{1-c^2} \right\rceil \lceil H_s \rceil B\right)\right)$*-head approximation algorithm. Moreover,* HeadApprox *runs in* $O(s\,n\,B\,h)$ *time for fixed* c.

Proof. Let $X' \in \mathbb{R}^{h \times w}$ with $X'_{i,j} = X^2_{i,j}$. We run HeadApprox with parameters X', k, B and c^2. Let Ω be the support returned by HeadApprox. Let $k' = \left\lceil \frac{4c^2}{1-c^2} \right\rceil k$ and $B' = \left\lceil \frac{4c^2}{1-c^2} \right\rceil \lceil H_s \rceil B$. Then according to Theorem 23 we have $\Omega \in \mathbb{M}_{k',B'}$. Moreover, we get the guarantee that $\|x'_\Omega\|_1 \geq c^2 \max_{\Omega \in \mathbb{M}_{k,B}} \|x'_\Omega\|_1$, which directly implies that $\|x_\Omega\| \geq c \max_{\Omega \in \mathbb{M}_{k,B}} \|x_\Omega\|$. □

A.3 Tail Approximations

In this section, we describe the tail approximation oracle introduced in Section 6 in greater detail. As before, recall that s is the per-column sparsity in the EMD-model, i.e., $s = k/w$. We first formally define the support induced by a set of paths.

Definition 25 (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 $P = \{p_1, \ldots, p_s\}$ 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 p_i are sorted vertically and $i \leq j$, then $(u,v) \in p_i$ and $(w,v) \in p_j$ implies $u < w$). Then the paths in P define a support*

$$\Omega_P = \{(u,v) \mid (u,v) \in p_i \text{ for some } i \in [s]\}. \tag{9}$$

Now, we introduce the property connecting paths and supports.

Theorem 26. *Let $X \in \mathbb{R}^{h \times w}$ be a signal matrix, k be a sparsity parameter and $\lambda \geq 0$. Let $P = \{p_1, \ldots, p_s\}$ 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*

$$c(f_P) = -\|X_{\Omega_P}\|_1 + \lambda \operatorname{EMD}(\Omega_P)\,. \tag{10}$$

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 $-\|X_{\Omega_P}\|_1$. 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 $\lambda \operatorname{EMD}(\Omega_P)$. □

We also formalize the connection between min-cost flows in $G_{X,k,\lambda}$ and good supports in X.

Lemma 27. *Let $G_{X,k,\lambda}$ be an EMD flow network and let f be a min-cost flow in $G_{X,k,\lambda}$. Then f can be decomposed into s disjoint paths $P = \{p_1, \ldots, p_s\}$ which do not intersect vertically. Moreover,*

$$\begin{aligned}\|X - X_{\Omega_P}\|_1 + \lambda \operatorname{EMD}(\Omega_P) = \min_{\Omega \in \mathbb{M}_{k,B}} &\|X - X_\Omega\|_1 \\ &+ \lambda \operatorname{EMD}(\Omega)\,.\end{aligned} \tag{11}$$

Proof. Note that $\|X - X_\Omega\|_1 = \|X\|_1 - \|X_\Omega\|_1$. Since $\|X\|_1$ does not depend on Ω, minimizing $\|X - X_\Omega\|_1 + \lambda \operatorname{EMD}(\Omega)$ with respect to Ω is equivalent to minimizing $-\|X_\Omega\|_1 + \lambda \operatorname{EMD}(\Omega)$. Further, all edges and nodes in $G_{X,k,\lambda}$ have capacity one, so f can be composed into disjoint paths P. Since $G_{X,k,\lambda}$ has integer capacities, the flow f is integral and therefore P contains exactly s paths. Moreover, the paths in P are not intersecting vertically: if p_i and p_j intersect vertically, we can relax the intersection to get a set of paths P' with smaller support EMD and 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 $c(f_Q) = -\|X_\Omega\|_1 + \lambda \operatorname{EMD}(\Omega)$. Since f is a min-cost flow, so $c(f) \leq c(f_Q)$. The statement of the theorem follows. □

The parameters d and δ for TailApprox quantify the acceptable tail approximation ratio (see Theorem 18). In the algorithm, we assume that MinCostFlow$(G_{X,k,\lambda})$ returns the support corresponding to a min-cost flow in $G_{X,k,\lambda}$. We now prove the main result for TailApprox: a *bicriterion*-approximation guarantee that allows us to use TailApprox as a tail approximation algorithm. We show that one of the following two cases occurs:

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

Case 2: We get a solution with bounded tail approximation error and support-EMD at most B.

Before we prove the main theorem, we show that TailApprox always returns the optimal result for signals $X \in \mathcal{M}_{k,B}$, i.e., X itself.

Lemma 28. *For any $X \in \mathcal{M}_{k,B}$,* TailApprox*(X, k, B, d, δ) returns X for any d and δ.*

Algorithm 4 Tail approximation algorithm

function TAILAPPROX(X, k, B, d, δ)
  $x_{\min} \leftarrow \min_{|X_{i,j}|>0}|X_{i,j}|$, $\varepsilon \leftarrow \frac{x_{\min}}{wh^2}\delta$
  if X is s-sparse in every column **then**
    $\lambda_0 \leftarrow \frac{x_{\min}}{2wh^2}$, $\Omega \leftarrow$ MINCOSTFLOW(G_{X,k,λ_0})
    if $\Omega \in \mathbb{M}_{k,B}$ **then**
      return X_Ω
  $\lambda_r \leftarrow 0$, $\lambda_l \leftarrow \|X\|_1$
  while $\lambda_l - \lambda_r > \varepsilon$ **do**
    $\lambda_m \leftarrow (\lambda_l + \lambda_r)/2$, $\Omega \leftarrow$ MINCOSTFLOW(G_{X,k,λ_m})
    if EMD(Ω) $\geq B$ and EMD(Ω) $\leq dB$ **then**
      return X_Ω
    if EMD(Ω) $> B$ **then**
      $\lambda_r \leftarrow \lambda_m$
    else
      $\lambda_l \leftarrow \lambda_m$
  $\Omega \leftarrow$ MINCOSTFLOW(G_{X,k,λ_l}), **return** X_Ω

Proof. Since $X \in \mathcal{M}_{k,B}$, every column of X is s-sparse. We show that the following call MINCOSTFLOW($G_{X,k,\lambda_{\min}}$) returns an $\Omega \in \mathbb{M}_{k,B}$ with $\text{supp}(X) \subseteq \Omega$. First, we prove that MINCOSTFLOW($G_{X,k,\lambda_{\min}}$) returns a support set covering all nonzero entries in X. As a result, $\text{supp}(X) \subseteq \Omega$. Let Γ be any s-column-sparse support set not covering all entries in X and let Δ be any s-column-sparse support set covering all entries in X. So $\|X - X_\Gamma\|_1 \geq x_{\min}$ and $\|X - X_\Delta\|_1 = 0$. Hence, we get:

$$\begin{aligned} -\|X_\Delta\|_1 + \lambda_0 \text{EMD}(\Delta) &= -\|X_\Delta\|_1 + \frac{x_{\min}}{2wh^2}\text{EMD}(\Delta) \\ &< -\|X_\Delta\|_1 + x_{\min} \leq -\|X_\Gamma\|_1 \\ &\leq -\|X_\Gamma\|_1 + \lambda_0 \text{EMD}(\Gamma)\,. \end{aligned}$$

Therefore, the cost of the flow corresponding to Δ is always less than the cost of the flow corresponding to Γ. Next, we show that among the support sets covering all nonzero entries in X, MINCOSTFLOW($G_{X,k,\lambda_{\min}}$) returns a support set with minimum support-EMD. Since $X \in \mathcal{M}_{k,B}$, there is a $\Gamma \in \mathbb{M}_{k,B}$ with $\|X_\Gamma\|_1 = \|X\|_1$. Moreover, Ω is the support returned by MINCOSTFLOW($G_{X,k,\lambda_{\min}}$), so we have $\|X_\Omega\|_1 = \|X\|_1$ and

$$-\|X_\Omega\|_1 + \lambda_0 \text{EMD}(\Omega) \leq -\|X_\Gamma\|_1 + \lambda_0 \text{EMD}(\Gamma)\,. \tag{12}$$

So EMD(Ω) $\leq$ EMD(Γ) $\leq B$. Since X_Ω is also s-sparse in each column, $\Omega \in \mathbb{M}_{k,B}$. □

In order to simplify the proof of the main theorem, we use the following shorthands: $\Omega_l =$ MINCOSTFLOW(G_{X,k,λ_l}), $\Omega_r =$ MINCOSTFLOW(G_{X,k,λ_r}), $b_l =$ EMD(Ω_l), $b_r =$ EMD(Ω_r), $t_l = \|X - X_{\Omega_l}\|_1$, $t_r = \|X - X_{\Omega_r}\|_1$. Moreover, we assume that $h = \Omega(\log w)$, i.e., the matrix X is not very wide and low.

Theorem 18. *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 EMD at most B, i.e., $OPT = \min_{\Gamma \in \mathbb{M}_{k,B}}\|X - X_\Gamma\|_1$.*

Then at least one of the following two guarantees holds for Ω: (i) either $B \leq \mathrm{EMD}(\Omega) \leq dB$ and $\|X - X_\Omega\|_1 \leq OPT$, or (ii) $\mathrm{EMD}(\Omega) \leq B$ and $\|X - X_\Omega\|_1 \leq (1+\delta)\frac{d}{d-1}OPT$. Moreover, TailApprox *runs in $O(snh \log \frac{\|X\|_1 n}{x_{\min}\delta})$ time, where $x_{\min} = \min_{|X_{i,j}|>0}|X_{i,j}|$.*

Proof. If $X \in \mathcal{M}_{k,B}$, TailApprox returns X, which means both guarantees hold (see Lemma 28). If $X \notin \mathcal{M}_{k,B}$ but X is s-sparse in each column, the following call MinCostFlow$(G_{X,k,\lambda_{\min}})$ returns an Ω covering all nonzero entries in X and using the minimum amount of support-EMD among all supports covering all nonzero entries in X (again, see Lemma 28). However, since $X \notin \mathcal{M}_{k,B}$, we have $\mathrm{EMD}(\Omega) > B$ and hence $\Omega \notin \mathbb{M}_{k,B}$. So TailApprox does not terminate early for $X \notin \mathcal{M}_{k,B}$. In the following, we assume that $X \notin \mathcal{M}_{k,B}$ and hence $OPT \geq x_{\min}$.

In the binary search, we maintain the invariant that $b_l \leq B$ and $b_r > B$. Note that this is true before the first iteration of the binary search due to our initial choices of λ_r and λ_l[8]. Moreover, our update rule maintains the invariant. We now consider the two ways of leaving the binary search. If we find an Ω with $\mathrm{EMD}(\Omega) \geq B$ and $\mathrm{EMD}(\Omega) \leq dB$, this also means $\|X - X_\Omega\|_1 \leq OPT$ due to convexity of the convex hull of supports. Hence the first guarantee in the theorem is satisfied. If $\lambda_l - \lambda_r \leq \varepsilon$, we return $\Omega = \Omega_l$ and hence the $\mathrm{EMD}(\Omega) \leq B$ part of the second guarantee is satisfied.

We now prove the bound on $\|X - X_\Omega\|_1 = t_l$. Figure 2 illustrates the geometry of the following argument. Let P_{OPT} be the point corresponding to a support with tail error OPT and minimum support-EMD, i.e., the optimal solution. Since the point (b_r, t_r) was the result of the corresponding MinCostFlow(G_{X,k,λ_r}), P_{OPT} has to lie above the line with slope $-\lambda_r$ through (b_r, t_r). Moreover, P_{OPT} has to have x-coordinate less than B. We can use these facts to establish the following bound on OPT:

$$OPT \geq t_r + \lambda_r(b_r - B)\,. \tag{13}$$

Let λ be the slope of the line through (t_r, b_r) and (t_l, b_l), i.e., $\lambda = \frac{t_r - t_l}{b_r - b_l}$. Then we have $\lambda_r \leq -\lambda \leq \lambda_l$. Together, with $\lambda_l - \lambda_r \leq \varepsilon$, this gives $\lambda_r \geq \frac{t_l - t_r}{b_r - b_l} - \varepsilon$. We now use this bound on λ_r to derive a bound on OPT:

$$\begin{aligned}
OPT &\geq t_r + \lambda_r(b_r - B) \\
&\geq t_r + (b_r - B)\frac{t_l - t_r}{b_r - b_l} - \varepsilon(b_r - B) \\
&\geq t_r + (b_r - B)\frac{t_l - t_r}{b_r} - \varepsilon(wh^2 - B) \\
&\geq t_l - \frac{B}{b_r}(t_l - t_r) - \varepsilon(wh^2 - B) \\
&\geq t_l - \frac{B}{dB}t_l - \frac{x_{\min}}{wh^2}\delta(wh^2 - B) \\
&\geq \frac{d-1}{d}t_l - x_{\min}\delta \\
&\geq \frac{d-1}{d}t_l - \delta OPT\,.
\end{aligned}$$

Hence,

$$t_l \leq (1+\delta)\frac{d}{d-1}OPT\,, \tag{14}$$

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

which shows that the second guarantee of the theorem is satisfied.

We now analyze the running time of TailApprox. 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 $O(wh\log(wh)+wh^2)$ time. The binary search takes at most $\log\frac{\|X\|_1}{\epsilon}=\log\frac{\|X\|_1 nh}{x_{\min}\delta}$ iterations. Using $n=wh$ gives the stated running time. □

Corollary 19. *Let $\delta>0$ and $d=1+\frac{1}{c^2/(1+\delta)-1}$. Then* TailApprox *is a $(c,(k,dB))$-tail approximation algorithm.*

Proof. Let $X'\in\mathbb{R}^{h\times w}$ with $X'_{i,j}=X^2_{i,j}$. We run TailApprox with parameters X',k,B,d and δ. Let Ω be the support returned by TailApprox. The tail approximation guarantee follows directly from Theorem 18. Note that we can not control which of the two guarantees the algorithm returns. However, in any case we have $\mathrm{EMD}(\Omega)\le dB$, so $\Omega\in\mathbb{M}_{k,dB}$. Moreover, note that $(1+\delta)\frac{d}{d-1}=c^2$. Rearranging, we get the guarantee that $\|x-x_\Omega\|_2\le c\min_{x^*\in\mathcal{M}_{k,B}}\|x-x^*\|_2$, which is precisely the claim made in Corollary 19. □

A.4 Proof of Theorem 20

Theorem 20. *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 $\|x-\widehat{x}\|_2\le C\|e\|_2$, for some constant C, from $m=O(k\log(\frac{B}{k}\log(\frac{k}{w})))$ measurements. Moreover, the recovery algorithm runs in time $O(snh\log\frac{\|x\|_2}{\|e\|_2}(B+d\log(\|x\|_2 n)))$ if x, A and e are specified with at most d bits of precision.*

Proof. We use AM-IHT, TailApprox and HeadApprox. The output $\widehat{x}$ of AM-IHT is the result of TailApprox with parameters k, B and c_T. As shown in Corollary 14, $c_T=1.5$ suffices for geometric convergence of AM-IHT. With this choice of c_T, TailApprox is a $(1.5,(k,2B))$-tail approximation algorithm (Corollary 19 and choosing $\delta=0.1$). Hence $\widehat{x}\in\mathcal{M}_{k,2B}$. We show that $m=O(k\log(\frac{B}{k}\log(\frac{k}{w})))$ suffices for A to have the (δ,t)-model-RIP for $t=f_H(p\oplus f_T(p))\oplus p\oplus f_T(p)$. Note that for the EMD-model, we have $p=(k,B)$ and $\mathcal{M}_{p\oplus q}\subseteq\mathcal{M}_{p+q}$, so we are interested in $t=f_H(p+f_T(p))+p+f_T(p)$. For $c_H=0.95$ (Corollary 14), HeadApprox is a $(0.95,(38k,38\lceil H_s\rceil B))$-head approximation oracle. So we get $t=(k',B')$ with $k'=\Theta(k)$ and $B'=\Theta(B\log s)$. From Theorem 8, we get

$$\begin{aligned}\log a_{k',B'}&=O\left(k'\log\frac{B'}{k'}\right)\\&=O\left(k\log\left(\frac{B}{k}\log(s)\right)\right)\\&=O\left(k\log\left(\frac{B}{k}\log\left(\frac{k}{w}\right)\right)\right).\end{aligned}$$

Combining this with fact 4 shows that $m=O(k\log(\frac{B}{k}\log(\frac{k}{w})))$ is sufficient for A to have the desired (δ,t)-model-RIP for fixed δ. Therefore, all assumptions in the analysis of AM-IHT are satisfied.

Using Corollary 14 with a sufficiently large t (e.g. $25 \log \frac{\|x\|_2}{\|e\|_2}$) gives the desired approximation error bound with $C = 152$. □

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 $m = O(k \log \frac{n}{k})$.

B Counterexample for IHT with tail approximation oracles

In order to present a simple explanation, we look at the standard k-sparse compressive sensing setting. Let $a \in \mathbb{R}^n$ and let $T'_k(x)$ be a tail approximation oracle with the following guarantee:

$$\left\|a - T'_k(a)\right\|_2 \leq c\|a - T_k(a)\|_2 \,, \tag{15}$$

where c is an arbitrary constant and T_k is an optimal projection oracle, i.e., returns a k-sparse vector with the k largest components of a. We now show that an "adversarial" tail approximation oracle with $T'_k(a) = 0$ satisfies this definition for inputs a occurring in the first iteration of IHT with high probability. This shows that IHT cannot make progress and consequently we cannot recover the signal.

Recall that IHT with tail approximation oracle T' iterates

$$x^{i+1} \leftarrow T'_k(x^i + \Phi^T(y - \Phi x^i)) \,, \tag{16}$$

which in the first iteration gives

$$x^1 \leftarrow T'_k(\Phi^T y) \,. \tag{17}$$

We now look at the case that 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 Φ with $(\delta, O(1))$-RIP for small δ, IHT needs to perfectly recover x from Φx. Matrices $\Phi \in \mathbb{R}^{m \times n}$ with $\Phi_{i,j} = \pm\sqrt{m}$ i.i.d. uniformly at random are known to have this RIP for $m = O(\log n)$ with high probability (so called Rademacher matrices). We prove that our adversarial tail approximation oracle satisfies the tail approximation guarantee for its input $a = \Phi^T \Phi e_1$ with high probability. Hence $x^1 = x^0 = 0$ and IHT cannot make progress. Intuitively, in spite of the RIP, the tail of a contains so much "noise" that the adversarial tail approximation oracle does not have to find a good sparse support for a and can get away with simply returning 0.

Consider the components of the vector $a \in \mathbb{R}^n$: a_i is the inner product of the first column of Φ with the i-th column of Φ. We have $a_1 = 1$ and $-1 \leq a_i \leq 1$ for $i \neq 1$. Hence $T_k(a) = e_1$ is an optimal projection and so $\|a - T_k(a)\|_2^2 = \|a\|_2^2 - 1$ We want to show that $\|a\|_2^2 \geq \frac{c^2}{c^2-1}$. This statement is equivalent to

$$\|a\|_2^2 \leq c^2(\|a\|_2^2 - 1) \,, \tag{18}$$

which then implies that T' satisfies the desired guarantee

$$\left\|a - T'_k(a)\right\|_2^2 \leq c^2\|a - T_k(a)\|_2^2 \,. \tag{19}$$

Note 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 ($p = 1/2$). We have $\mathbb{E}[a_i^2] = \frac{1}{m}$. 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

$$\|a\|_2^2 = 1 + \sum_{i=2}^{n} a_i^2 \geq \frac{c^2}{c^2 - 1} \tag{20}$$

with high probability for sufficiently large n.

C Figures

$$X = \begin{bmatrix} 1 & 3 \\ 0 & -1 \\ 2 & 1 \end{bmatrix}$$

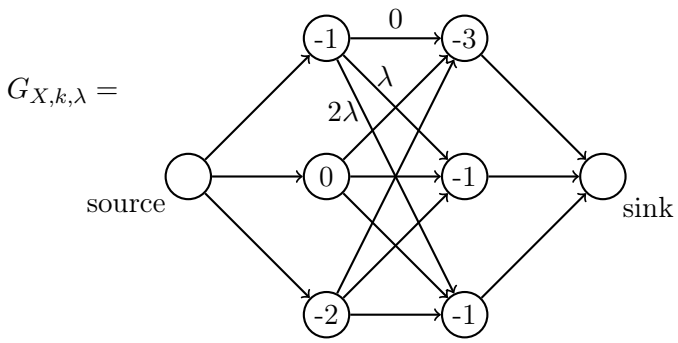


Figure (1): A signal X with the corresponding flow network $G_{X,k,\lambda}$. The node costs are the negative absolute values of the corresponding signal components. The numbers on edges indicate the edge costs (most edge costs are omitted for clarity). All capacities in the flow network are 1. The edge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

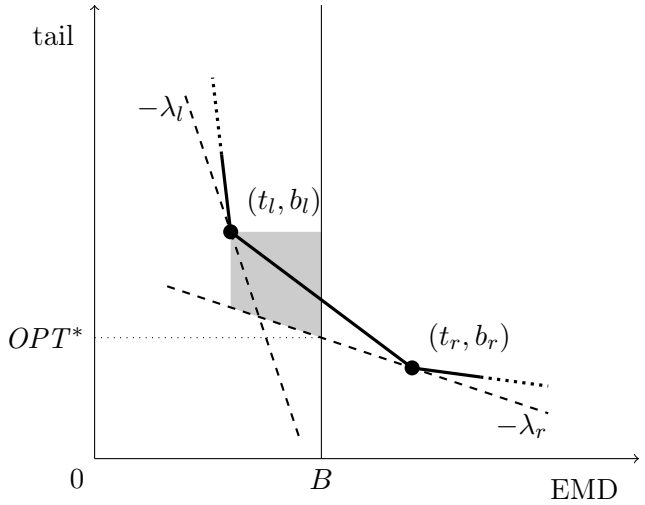


Figure (2): The local region of the convex hull during the binary search. The point (t_l, b_l) corresponds to MinCostFlow(G_{X,k,λ_l}) and (t_r, b_r) corresponds to MinCostFlow(G_{X,k,λ_r}). All support points between the two points have to lie above the dashed lines with slopes $-\lambda_l$ and $-\lambda_r$. We also use the fact that the optimal support has to have a x-coordinate between b_l and B and a y-coordinate below t_l. In the proof of Theorem 18 we use only the line corresponding to λ_r, which leaves the gray area. As a result, OPT^* is a lower bound on OPT.

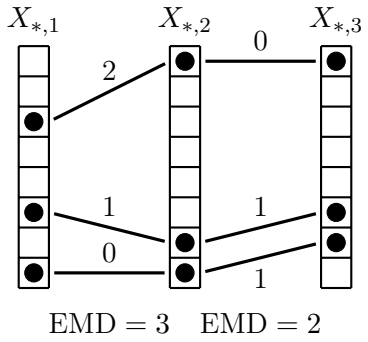
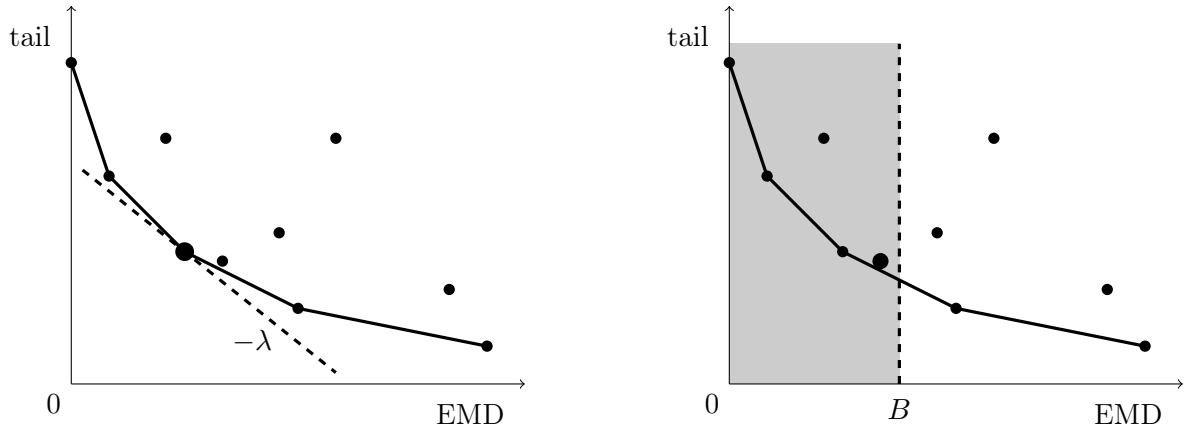


Figure (3): The support EMD for a matrix with three columns and eight rows. The circles stand for supported elements in the columns. The lines indicate the matching between the supported elements and the corresponding EMD cost. The total support EMD is $\mathrm{EMD}(\mathrm{supp}(X)) = 2 + 3 = 5$.

$$X = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 1 & 2 \\ 4 & 2 & 0 \end{bmatrix} \qquad X^* = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 2 \\ 4 & 2 & 0 \end{bmatrix}$$

Figure (4): 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.



(a) Result of a min-cost flow run.

(b) Optimal solution

Figure (5): Each point corresponds to a support Ω for the matrix X. The x-coordinate of Ω is EMD(Ω) and the y-coordinate is $\|X - X_\Omega\|_1$. Finding min-cost flows in $G_{X,k,\lambda}$ allows us to find points on the convex hull of support points.
The dashed line in figure (a) illustrates the result of MinCostFlow($G_{X,k,\lambda}$), which is also the slope of the line. The point found is the first point we hit when we move a line with slope $-\lambda$ from the origin upwards (the larger dot in the figure).
For a given EMD budget B, we want to find the support with the smallest tail. The shaded region in figure (b) indicates the region where supports with support-EMD at most B lie. We want to find the point in this region with minimum y-coordinate. Note that this point (the larger dot in the figure) does not necessarily lie on the convex hull.