DIMACS Technical Report 2005-40

Combinatorial Algorithms for Compressed Sensing

by

Graham Cormode[1]
Bell Labs
cormode@bell-labs.com

S. Muthukrishnan[2,3]
Rutgers University
muthu@cs.rutgers.edu

DIMACS is a collaborative project of Rutgers University, Princeton University, AT&T Labs–Research, Bell Labs, NEC Laboratories America and Telcordia Technologies, as well as affiliate members Avaya Labs, HP Labs, IBM Research, Microsoft Research, Stevens Institute of Technology, Georgia Institute of Technology and Rensselaer Polytechnic Institute. DIMACS was founded as an NSF Science and Technology Center.

ABSTRACT

In sparse approximation theory, the fundamental problem is to reconstruct a signal $\mathbf{A} \in \mathbb{R}^n$ from linear measurements $\langle \mathbf{A}, \psi_i \rangle$ with respect to a dictionary of ψ_i's. Recently, there is focus on the novel direction of *Compressed Sensing* [9] where the reconstruction can be done with very few—$O(k \log n)$—linear measurements over a modified dictionary if the signal is *compressible*, that is, its information is concentrated in k coefficients with the original dictionary. In particular, these results [9, 4, 22] prove that there exists a single $O(k \log n) \times n$ measurement matrix such that any such signal can be reconstructed from these measurements, with error at most $O(1)$ times the worst case error for the class of such signals. Compressed sensing has generated tremendous excitement both because of the sophisticated underlying Mathematics and because of its potential applications.

In this paper, we address outstanding open problems in Compressed Sensing. Our main result is an explicit construction of a non-adaptive measurement matrix and the corresponding reconstruction algorithm so that with a number of measurements polynomial in $k, \log n, 1/\varepsilon$, we can reconstruct any compressible signal. This is the first known polynomial time explicit construction of any such measurement matrix. In addition, our result improves the error guarantee from $O(1)$ to $1 + \varepsilon$ and improves the reconstruction time from $\text{poly}(n)$ to $\text{poly}(k \log n)$.

Our second result is a randomized construction of $O(k \,\text{polylog}(n))$ measurements that work for each signal with high probability and gives per-instance approximation guarantees rather than over the class of all signals. Previous work on Compressed Sensing does not provide such per-instance approximation guarantees; our result improves the best known number of measurements known from prior work in other areas including Learning Theory [20], Streaming algorithms [11, 12, 6] and Complexity Theory [1] for this case.

Our approach is combinatorial. In particular, we use two parallel sets of group tests, one to filter and the other to certify and estimate; the resulting algorithms are quite simple to implement.

1 Introduction

We study a modern twist to a fundamental problem in sparse approximation theory, called *Compressed Sensing* which has been proposed recently in the Mathematics community.

Sparse Approximation Theory Background. The *dictionary* Ψ denotes an orthonormal basis for $\mathbb{R}^n$, i.e. Ψ is a set of n real-valued vectors ψ_i each of dimension n and $\psi_i \perp \psi_j$. The *standard basis* is the traditional coordinate system for n dimensions, namely, for $i = 1, \ldots, n$, the vector $\psi_i = [\psi_{i,j}]$ where $\psi_{i,j} = 1$ iff $i = j$.[1] A *signal* vector $\mathbf{A}$ in $\mathbb{R}^n$ is transformed by this dictionary into a vector of *coefficients* $\theta(\mathbf{A})$ formed by inner products between $\mathbf{A}$ and vectors from Ψ. That is, $\theta_i(\mathbf{A}) = \langle \mathbf{A}, \psi_i \rangle$ and $\mathbf{A} = \sum_i \theta_i(\mathbf{A})\psi_i$ by the orthonormality of Ψ.[2] From now on (for convenience of reference only), we reorder the vectors in the dictionary so $|\theta_1| \geq |\theta_2| \geq \ldots \geq |\theta_n|$.

In the area of sparse approximation theory [8], one seeks representations of $\mathbf{A}$ that are *sparse*, i.e., use few coefficients. Formally, $\mathbf{R} = \sum_{i \in K} \theta_i \psi_i$, for some set K of coefficients, $|K| = k \ll n$. Clearly, $\mathbf{R}(\mathbf{A})$ cannot exactly equal the signal $\mathbf{A}$ for all signals. The error is typically taken as $\|\mathbf{R} - \mathbf{A}\|_2^2 = \sum_i (\mathbf{R}_i - \mathbf{A}_i)^2$. By the classical Parseval's equality, this is equivalently $\|\theta(\mathbf{A}) - \theta(\mathbf{R})\|_2^2$. The optimal k representation of $\mathbf{A}$ under Ψ, $\mathbf{R}^k_{\text{opt}}$, therefore takes k coefficients with the largest $|\theta_i|$'s. The error then is $\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}\|_2^2 = \sum_{i=k+1}^n \theta_i^2$. This is the error in representing the signal $\mathbf{A}$ in a compressed form using k coefficients from Ψ.

In any application (say audio signal processing), one has a "class" of input signals ($\mathbf{A}$'s) (e.g.m sinusoidal waveforms comprising the audio signal), one chooses an appropriate dictionary Ψ (say discrete Fourier) so that most of the signals are "compressible" using that dictionary, and represents the signal using the adequate number ($k \ll n$) of coefficients $(\theta_1, \ldots, \theta_k)$. There are different notions of a signal being compressible in a dictionary. In the past, e.g., in audio applications, researchers focused on the *α-exponentially decaying case* where the coefficients decay faster than any polynomial. That is, for some α, $|\theta_i| = \Theta(2^{-\alpha i})$, for all i. More recently, there is focus on the *p-Compressible case.* Specifically the coefficients have a power-law decay: for some $p \in (0, 1)$, and for all i, $|\theta_i| = \Theta(i^{-1/p})$. Consequently, $\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}\|_2^2 \leq C_p k^{1-2/p}$ for some constant C_p.

Study of sparse approximation problems involves the art of identifying suitable Ψ so the signals from an application are compressible, and studying their mathematical properties. This is a mature area of Mathematics with highly successful applications to signal processing, communication theory and compression [8].

Compressed Sensing. Recently, Donoho posed a fundamental question [9]. Since most of the information in the signal is contained in only a few coefficients and the rest of the signal is not needed for the applications, can one directly determine (acquire) only the relevant coefficients without reading (measuring) each of the coefficients? In a series of papers over the past year, the following result has emerged.

[1] Examples of other basis are *discrete Fourier* where $\psi_{i,j} = \frac{1}{\sqrt{n}} \exp(-2\pi\sqrt{-1}ij/n)$; and *Haar wavelet* where every ψ_i is a scaled and shifted copy of the same step like function. By applying an appropriate rotation to the basis and signal vectors, our problem can be thought of in the standard basis only.

[2] We refer to θ_i where $\mathbf{A}$ is implicitly clear.

Theorem 1. *[9, 4, 22] There exists a non-adaptive set V of $O(k\log(n/k))$ vectors in $\mathbb{R}^n$ which can be constructed once and for all from the standard basis. Then, for fixed $p \in (0,1)$ and any p-compressible signal $\mathbf{A}$ in the standard basis, given only* measurements $\langle \mathbf{A}, v_i\rangle$, *$v_i \in V$, a representation $\mathbf{R}$ can be determined in time polynomial in n such that $\|\mathbf{A}-\mathbf{R}\|_2^2 = O(k^{1-2/p})$.*

There are several important points to note. First, since the worst case error for a p-compressible signal is $C_p k^{1-2/p}$, the representation above is *optimal, up to constant factors* for the class of all p-compressible signals, for a fixed p. Second, even if the signal consisted of precisely k nonzero coefficients $\theta_{i_1}, \ldots, \theta_{i_k}$, one needs k measurements $\langle \mathbf{A}, \psi_{i_j}\rangle$ for $j \in [1,k]$; hence, the set V of measurements is only a $\log(n/k)$ factor larger than the naive lower bound of measurements needed. Third, the proof shows existence of V by showing that a random set of V vectors will satisfy the theorem with nonzero probability. The proof immediately gives a *Monte Carlo* randomized algorithm by using such a random V.

This result has generated great excitement. A slew of results have strengthened the result in different aspects [9, 24, 4, 22]; found interesting applications including MR imaging [23] wireless communication [22] and generated implementations [21]; found mathematical applications to coding and information theory [3]; and extended the results to noisy and distributed settings [2]. The excitement arises for two main reasons. First, there is deep mathematics underlying the results, with interpretations in terms of high dimensional geometry [22], uncertainty principles [4], and linear algebra [9]. Second, there are serious applications—for example, in going from analog to digital representation of the signals, existing hardware chips can execute measurements $\langle \mathbf{A}, v_i\rangle$ extremely efficiently, so performing $O(k\log(n/k))$ measurements is significantly more efficient than measuring each component of the signal (hence "compressed sensing"). The results have inspired a number of recent workshops, meetings and talks [23, 15, 18].

Outstanding Problems and Our Results. There are several outstanding questions in Compressed Sensing.

The most fundamental issue is to explicitly construct the non-adaptive measurement set of vectors V (or equivalently, a *transformation matrix* T in which $T[i,j] = v_i[j]$) in the theorem. The existing results first show that if T satisfies certain conditions, then the theorem holds; then, they show that T chosen from an appropriate random distribution suffices. The necessary conditions are quite sophisticated, such as computing the eigenvalues of every $O(k\log n)$ square submatrix of T [9], and testing that each such submatrix is an isometry, behaving like an orthonormal system [4]. No explicit construction is known to produce T's with these properties! Instead, algorithms for Compressed Sensing choose a random T, and assume that the conditions are met. Thus, these are Monte Carlo algorithms, with some probability of failure. This is a serious drawback for Compressed Sensing applications motivated by hardware implementations which will sense many, many signals over time. So it is highly desirable that there be an explicit construction of T suitable for Compressed Sensing. A natural approach is to take a random T and test whether it satisfies the necessary conditions. However, this is much too expensive, taking time at least $\Omega(n^{k\log n})$.

There are several other outstanding questions. For example, the time to obtain a representation from the measurements is significantly *superlinear* in n (it typically involves solving a Linear Program [9, 4, 22]). Since we make a small number of measurements, it is much preferable to find algorithms with running time polynomial in the number of measurements and hence, sublinear in n. Lastly, the guarantee given by the above theorem is not relative to the best possible for the given signal (i.e., *per-instance*), but to the *worst case* over the whole class of p-compressible signals. Clearly per-instance error guarantees are preferable.

We address these outstanding questions and present the first known completely explicit algorithms for Compressed Sensing. Our approach is combinatorial, and yields a number of technical improvements such as sublinear time reconstruction, and tolerance to error. Our main results are twofold.

1. We present a deterministic algorithm that in time polynomial in k and n constructs a non-adaptive transformation matrix T of number of rows polynomial in $k \log n$, and present an associated reconstruction algorithm in the spirit of Theorem 1. More specifically, our algorithm outputs a representation $\mathbf{R}$ such that $\|\mathbf{R} - \mathbf{A}\|_2^2 < \|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 + \varepsilon C_p k^{1-2/p}$. This is the first explicit construction known for this problem in polynomial time.

In addition, this result leads to the following improvements: (a) the reconstruction time is subquadratic in the number of measurements (and hence sublinear in n), (b) the overall error is optimal up to $1+\varepsilon$ of the worst case error $C_p k^{1-2/p}$ over the class of p-compressible signals, improving the $O(1)$ factor in prior results, and (c) the approach applies to other cases of compressible signals with tighter bounds. For the exponentially decaying family, the size of T is only $O(k^2 \operatorname{polylog}(n))$. The algorithms are simple and easy to implement, without linear programming and without running into precision-issues inherent in the choice of Gaussian random T in prior methods.

2. We address the issue of obtaining per-instance guarantees for each signal. We present a randomized algorithm that on any given $\mathbf{A}$, produces a T with $O(k \operatorname{polylog}(n))$ rows such that in time linear in $O(k \operatorname{polylog}(n))$, we can reconstruct a $\mathbf{R}$ with $\|\mathbf{A}-\mathbf{R}\|_2^2 \leq (1+\varepsilon)\|\mathbf{A}-\mathbf{R}^k_{\text{opt}}\|_2^2$, with probability at least $1 - \frac{1}{n^{O(1)}}$.

Notice crucially that this second result does not produce a T that works for all p-compressible signals, merely, that on any given signal $\mathbf{A}$, we can produce a good $\mathbf{R}$ with high probability. In this regime, which is quite different from the regime in earlier papers on Compressed Sensing where a fixed T works for all p-compressible signals, many results in the Computer Science literature apply, in particular, from learning theory [20], streaming algorithms [12, 11] and complexity theory [1]. These results can be thought of as producing a T with $O(k^{2+O(1)} \operatorname{polylog}(n))$ rows which is improved by our result here. An exception is the result in [13] which works by sampling (that is, finding $\langle \mathbf{A}, v_i \rangle$ where $v_{i,j} = 1$ for some j and is 0 elsewhere) for the Fourier basis, but can be thought of as solving our problem using $O(k \operatorname{polylog}(1/\varepsilon, \log n, \log \|\mathbf{A}\|)$ measurements. Our result improves [13] in the term $\operatorname{polylog}(\|\mathbf{A}\|_2)$ which governs the number of iterations in [13]. Finally, we extend to the case when the measurements are noisy—an important practical concern articulated in [14]—and obtain first known results that give per-instance approximation results.

Technical Overview. The intuitive way to think about these problems is to consider combinatorial group testing problems. We have a set $U = [n]$ of items and a set D of *distinguished* items, $|D| \leq k$. We identify the items in D by performing group tests on subsets $S_i \subseteq U$ whose output is 1 or 0, revealing whether that subset contains one or more distinguished items, that is $|S_i \cap D| \geq 1$. There exist collections of $O((k \log n)^2)$ nonadaptive tests which identify each of the distinguished items precisely.

There is a strong connection between this problem and Compressed Sensing. We can treat θ_i's as items and the largest (in magnitude) k as the members of D. Each test set S_i can be written as its characteristic vector χ_{S_i} of n dimensions. A difficulty arises in interpreting the outcome of $\langle \mathbf{A}, \chi_{S_i} \rangle$. The discussion so far has been entirely combinatorial, but the outcome of this linear-algebraic operation of inner product must be interpreted as a binary outcome to apply standard combinatorial group testing methods. In general, there is no direct connection between $\langle \mathbf{A}, \chi_{S_i} \rangle$ and presence or absence of the first k coefficients in S_i when the signal is from the p-compressible class. This is also the reason that prior work on this problem has delved into the linear-algebraic and geometric structure of the problem.

Our approach here is combinatorial. Our first results show that one can focus attention on some $k' > k$ coefficients, in order to meet our error guarantees. Then, we show that separating the k' coefficients using group testing methods serves as a filter and subsequently, using a different set of group tests serves to certify and estimate the largest k coefficients in magnitude. This use of two parallel sets of group tests is novel. For the second set of results, combinatorial group testing has been applied previously in Learning Theory [20], Streaming Algorithms [11, 12, 6] and Complexity Theory [1]. Here, our contribution is to adapt the approach from our first set of results and provide a tighter analysis of the error in terms of $\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2$ rather than in terms of $\|\mathbf{A}\|_2$ as is more typical.

Note. A preliminary version of this paper appeared as a tech report [7] and is superseded by the results here.

1.1 Preliminaries

Definition 1. *A collection $\mathcal{S}$ of l subsets of $\{1 \ldots n\}$ is called* k-selective *if for any X such that $X \subset \{1 \ldots n\}$ and $|X| \leq k$, there exists $S_i \in \mathcal{S}$ such that $|S_i \cap X| = 1$, i.e. there is a member of X which occurs separated from all other members of X in some S_i.*

Definition 2. *A collection $\mathcal{S}$ of m subsets of $\{1 \ldots n\}$ is called* k-strongly selective *if for any X with $|X| \leq k$, and for all $x \in X$ there exists $S_i \in \mathcal{S}$ such that $S_i \cap X = \{x\}$, i.e. every member of X occurs separated from all other members of X in some S_i.*

We note that k-strongly selectivity is a stronger condition than k-selectivity, and implies it. Explicit constructions of both collections of sets are known for arbitrary k and n. Strongly selective sets are used heavily in group testing [10], and can be constructed using superimposed codes [19] with $m = O((k \log n)^2)$. Indyk provided explicit constructions of k-selective collections of size $l = O(k \log^{O(1)} n)$, where the power depends on the degree bounds of constructions of disperser graphs [16]. Probabilistic constructions are also possible [5] of

near-optimal size $O(k \log(n/k))$, which yield a more expensive Las Vegas-style algorithm for constructing such a set in $O(n^k \operatorname{poly}(k \log n))$: for each of the $\binom{n}{k}$ choices of X, verify that the required property holds.

We will also make use of the Hamming code matrix H_n, which is the $\lceil 1 + \log_2 n \rceil$ matrix whose ith column is 1 followed by the binary representation of i. We will combine matrices together to get larger matrices by (a) concatenating the rows of N to M and get matrix denoted $M \bigcup N$, or (b) a Tensor product-like operation we denote $\otimes$, defined as follows:

Definition 3. *Given matrices V and W of dimension $v \times n$ and $w \times n$ respectively, define the matrix $(V \otimes W)$ of dimension $vw \times n$ as $(V \otimes W)_{iv+l,j} = V_{i,j} W_{l,j}$.*

2 Non-adaptive Constructions

We must describe the construction of a set of m (row) vectors Ψ' that will allow us to recover sufficient information to identify a good set of coefficients. We treat Ψ' as an $m \times n$ matrix whose ith row is Ψ'_i. When given the vector of measurements $\Psi'\mathbf{A}$ we must find an approximate representation of $\mathbf{A}$. Ψ' is a function of Ψ, and more strongly (as is standard in compressed sensing) we only consider matrices Ψ' that can be written as a linear combination of vectors from the dictionary Ψ, i.e., $\Psi' = T\Psi$, for some $m \times n$ transform matrix T. Thus $\Psi'\mathbf{A} = T(\Psi\mathbf{A}) = T\theta$. Recall that the best representation under Ψ using k coefficients is given by picking k largest coefficients from θ. We use T to let us estimate k large coefficients from θ, and use these to represent $\mathbf{A}$; we show that the error in this representation can be tightly bounded.

Observe that we could trivially use the identity matrix I as our transform matrix T. From this we would have $T\theta = \theta$, and so could recover $\mathbf{A}$ exactly. However, our goal is to use a transform matrix that is much smaller than the n rows of I, preferably polynomial in k and $\log n$. In general, the only way to achieve exact recovery of the optimal representation is to take a linear number of measurements:

Lemma 1. *Any deterministic construction which returns k coefficients and guarantees error exactly $\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2$ requires $\Theta(n)$ measurements.*

2.1 p-compressible signals

In the p-compressible case the coefficients (sorted by magnitude) obey $|\theta_i| = \Theta(i^{-1/p})$ for appropriate scaling constants and some parameter p. Previous work has focused on the cases $0 < p < 1$ [4, 9]. Integrating shows that $\sum_{i=k+1}^{n} \theta_i^2 = \|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 = O(k^{1-2/p})$. Our results, like those of [4, 9], are stated with respect to the error due to the worst case over all signals in the class, which we denote $\|\mathbf{C}^k_{\text{opt}}\|_2 = O(k^{1-2/p})$. For any signal that is p-compressible with fixed p and C_p it follows that $\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2 \leq \|\mathbf{C}^k_{\text{opt}}\|_2$.

Our transform collects information based on two collections of strongly separating sets. The first ensures that sufficient separation occurs, allowing all large coefficients to be recovered. The second allows accurate estimates of the weight of each coefficient to be made.

Transform Definition. We define our transform matrix as follows. Let $k' = c'(k\varepsilon^{-p})^{1/(1-p)^2}$ and $k'' = c''(k'\frac{\log n}{\log k'})^2$ for appropriately chosen c' and c''. Let $\mathcal{S}$ be a k'-strongly separating collection of sets (so that the number of sets in the collection is k''), and write T_1 as the matrix formed by the concatenation of χ_{S_i} for all S_i in $\mathcal{S}$. Similarly, let $\mathcal{R}$ be a k''-strongly separating collection of sets, and write T_2 as its characteristic matrix. We form our transform matrix T_p as $(T_1 \otimes H) \bigcup T_2$.

The intuition is that rather than ensuring separation for just the k largest coefficients, we will guarantee separation for the top-k' coefficients, where k' is chosen so that the remaining coefficients are so small that even if taken all together, the error introduced to the estimation of any coefficient is still within our allowable error bounds.

Reconstruction Algorithm. Our algorithm for recovering a representation from the results of the measurements $T_p\Psi\mathbf{A}$ is as follows: for each set of $\lceil 1 + \log n \rceil$ measurements due to $S_i \otimes H$, we recover $x_0 \dots x_{\lceil \log n \rceil} = (S_i \otimes H)\Psi\mathbf{A}$, and decode an identifier j_i as

$$j_i = \sum_{b=1}^{\log n} 2^{b-1} \frac{|x_b| - \min\{|x_b|, |x_0 - x_b|\}}{\max\{|x_b|, |x_0 - x_b|\} - \min\{|x_b|, |x_0 - x_b|\}}.$$

This generates a set of coefficients $J = \{j_1, j_2 \dots j_{k''}\}$. We then use the measurements due to T_2 to estimate the weight of each coefficient named in J: for each $j \in J$, we set $\hat{\theta}_j = \chi_{R_i}\Psi\mathbf{A}$ for $J \cap R_i = \{j\}$. The strong separation properties of $\mathcal{R}$ ensure that there will be at least one such R_i, and if there is more than one, then we can pick one arbitrarily. Our output is the set of k pairs $(j, \hat{\theta}_j)$ with the k largest values of $|\hat{\theta}_j|$.

Lemma 2. *Let K' denote the set of the k' largest coefficients.*

1. $\forall j \in K' :\ \theta_j^2 \geq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2 \Rightarrow j \in J.$
2. $\forall j \in J :\ |\hat{\theta}_j - \theta_j| \leq \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2.$

Proof. Observe that the square of the (absolute) sums of coefficients after removing the top k' is $(\sum_{i=k'+1}^{n} |\theta_i|)^2 = O(k'^{2-2/p})$. Over the whole class of p-compressible signals, this is bounded by $O(k'^{2-2/p}/k^{1-2/p})\|\mathbf{C}_{\text{opt}}^k\|_2^2$. Substituting in $k' \geq C(k\varepsilon^{-p})^{1/(1-p)}$ for an appropriately chosen constant C ensures $(\sum_{i=k'+1}^{n} |\theta_i|)^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$. Now consider θ_j that satisfies the condition in the lemma.

Although K' is unknown, we can be sure that, since $\mathcal{R}$ is k'-strongly separating, there is at least one set R_i such that $K' \cap R_i = \{j\}$. Consider the vector of measurements involving this set, $x = (\chi_{R_i} \oplus H)\Psi\mathbf{A}$. When $H_{j,b} = 1$, $|x_b| \geq |\theta_j| - \sum_{l \neq j \in R_i} H_{l,b}|\theta_l|$ and $|x_0 - x_b| \leq \sum_{l \neq j \in R_i}(1 - H_{l,b})|\theta_l|$. Since $\theta_j^2 > \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$, and $\sum_{l \neq j \in R_i} \theta_l \leq \sum_{l=k'+1}^{n} |\theta_l| \leq \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2$, we have $|\theta_j| > \sum_{l \neq j \in R_i} H_{l,b}|\theta_l| + (1 - H_{l,b})|\theta_l|$. Hence $\min\{|x_b|, |x_0 - x_b|\} = |x_0 - x_b|$, and $\max\{x_b|, |x_0 - x_b|\} = |x_b|$. Thus

$$\frac{|x_b| - \min\{|x_b|, |x_0 - x_b|\}}{\max\{|x_b|, |x_0 - x_b|\} - \min\{|x_b|, |x_0 - x_b|\}} = 1 = H_{j,b}.$$

Symmetrically, the results are reversed when $H_{j,b} = 0$, where

$$\frac{|x_b| - \min\{|x_b|, |x_0 - x_b|\}}{\max\{|x_b|, |x_0 - x_b|\} - \min\{|x_b|, |x_0 - x_b|\}} = 0 = H_{j,b}.$$

Thus the decoded identifier $j_i = \sum_{b=1}^{\log n} 2^{b-1} H_{j,b} = j$ and so $j \in J$, showing (1).

For (2), observe that $|J| \leq k''$, since each $R_i \in \mathcal{R}$ generates at most one $j \in J$. Hence, we can guarantee for each $j \in J$ there is at least one S_i such that $J \cap S_i = j$. We chose our k' to be sufficiently large that we can identify the $(k\varepsilon^{-p})^{1/1-p}$ largest coefficients; since J contains the $(k\varepsilon^{-p})^{1/1-p}$ largest coefficients, we can be sure that $|\hat{\theta}_j - \theta_j| = |\chi_{R_i}\Psi\mathbf{A} - \theta_j| = |\sum_{l \in R_i, l \neq j} \theta_l| \leq \sum_{l=(k\varepsilon^{-p})^{1/1-p}+1}^{n} |\theta_l| \leq \frac{\varepsilon}{5\sqrt{k}} \|\mathbf{C}_{\text{opt}}^k\|_2$. □

Lemma 3 (Reconstruction accuracy)**.** *Given $\hat{\theta}(\mathbf{A}) = \{\hat{\theta}_i(\mathbf{A})\}$ such that $(\hat{\theta}_i - \theta_i)^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$ if $\theta_i^2 \geq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$, picking the k largest coefficients from $\hat{\theta}(\mathbf{A})$ gives an error $\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2^2$ k-term representation of $\mathbf{A}$.*

Proof. As stated in the introduction, the error from picking the k largest coefficients exactly is $\|\theta(\mathbf{A}) - \theta(\mathbf{R}_{\text{opt}}^k)\|_2^2 = \sum_{i=k+1}^{n} \theta_i^2$ (where we index the θ_is in decreasing order of magnitude). We will write $\hat{\phi}_i$ for the ith largest approximate coefficient, and ϕ_i for its exact value. Let $\pi(i)$ denote the mapping such that $\phi_i = \theta_{\pi(i)}$, and let $\sigma(i)$ denote a bijection satisfying $\sigma(i) = j \Rightarrow (i > k \wedge \pi(i) \leq k \wedge j \leq k \wedge \pi(j) > k)$.

Picking the k largest approximate coefficients has energy error

$$\begin{aligned}
\|\mathbf{R} - \mathbf{A}\|_2^2 &= \sum_{i=1}^{k} (\phi_i - \hat{\phi}_i)^2 + \sum_{i=k+1}^{n} \phi_i^2 \\
&= \sum_{i \leq k} (\phi_i - \hat{\phi}_i)^2 + \sum_{i>k, \pi(i)>k} \phi_i^2 + \sum_{i>k, \pi(i)\leq k} \phi_i^2 \\
&\leq \sum_{i \leq k} \frac{\varepsilon^2}{25k} \|\mathbf{C}_{\text{opt}}^k\|_2^2 + \sum_{i>k, \pi(i)>k} \phi_i^2 + \sum_{i>k, \pi(i)\leq k} \phi_i^2
\end{aligned}$$

Consider i such that $i > k$ but $\pi(i) \leq k$: this corresponds to a coefficient that belongs in the top-k but whose estimate leads us to not choose it. Then either $\phi_i^2 \leq \frac{\varepsilon^2}{2k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$, i.e. the top-$k$ coefficient is small compared to the optimal error, or else our estimate of $\phi_{\sigma(i)}^2$ was too high. In this case $\hat{\phi}_i^2 < \hat{\phi}_{\sigma(i)}^2$ but $\phi_{\sigma(i)}^2 \leq \phi_i^2$. Assuming this, we can write

$$\begin{aligned}
\phi_i^2 - \phi_{\sigma(i)}^2 &= (\phi_i + \phi_{\sigma(i)})(\phi_i - \phi_{\sigma(i)}) \\
&= (|\phi_i| + |\phi_{\sigma(i)}|)(|\phi_i| - |\phi_{\sigma(i)}|) \\
&= (2|\phi_{\sigma(i)}| + |\phi_i| - |\phi_{\sigma(i)}|)(|\phi_i + \hat{\phi}_i - \hat{\phi}_i| - |\phi_{\sigma(i)} + \hat{\phi}_{\sigma(i)} - \hat{\phi}_{\sigma(i)}|) \\
&\leq (2|\phi_{\sigma(i)}| + |\phi_i| - |\phi_{\sigma(i)}|)(|\phi_i - \hat{\phi}_i| + |\phi_{\sigma(i)} - \hat{\phi}_{\sigma(i)}| + |\hat{\phi}_i| - |\hat{\phi}_{\sigma(i)}|) \\
&\leq (2|\phi_{\sigma(i)}| + \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2)(\frac{2\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2)
\end{aligned}$$

In the case that $\phi_i^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$ we can immediately write

$$\phi_i^2 - \phi_{\sigma(i)}^2 \leq \phi_i^2 \leq \frac{\varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2}{5\sqrt{k}} \cdot \frac{\varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2}{5\sqrt{k}} \leq (2|\phi_{\sigma(i)}| + \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2)(\frac{2\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2)$$

Substituting this bound into the expression above, we use the facts that $\sum_{j=1}^k |a_j| \leq \sqrt{k}(\sum_{j=1}^k a_j^2)^{1/2}$ and $\sum_{i>k,\pi(i)\leq k} \phi_{\sigma_i}^2 = \sum_{j\leq k,\pi(j)>k} \phi_j^2$, to obtain a bound on $\|\mathbf{R}-\mathbf{A}\|_2^2$ as

$$\sum_{i\leq k,\pi(i)\leq k} \frac{\varepsilon}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2 + \sum_{i>k,\pi(i)>k} \phi_i^2 + \sum_{i>k,\pi(i)\leq k} (\phi_{\sigma(i)}^2 + (2|\phi_{\sigma(i)}| + \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2)(\frac{2\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2))$$
$$\leq \tfrac{\varepsilon}{25}\|\mathbf{C}_{\text{opt}}^k\|_2^2 + (2\sqrt{k} + \tfrac{\varepsilon\sqrt{k}}{5})\tfrac{2\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2^2 + \textstyle\sum_{\pi(i)>k} \phi_i^2$$
$$\leq \tfrac{23\varepsilon}{25}\|\mathbf{C}_{\text{opt}}^k\|_2^2 + \textstyle\sum_{i>k} \theta_i^2 < \|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2^2$$ □

Theorem 2. *We can construct a set of $O((k\varepsilon^p)^{4/(1-p)^2}\log^4 n)$ measurements in time polynomial in k and n. For any p-compressible signal $\mathbf{A}$, from these measurements of $\mathbf{A}$, we can return a representation $\mathbf{R}$ for $\mathbf{A}$ of at most k coefficients $\hat{\theta}$ under Ψ such that $\|\hat{\theta}-\theta\|_2^2 = \|\mathbf{R}-\mathbf{A}\|_2^2 < \|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2^2$. The time required to produce the coefficients from the measurements is $O((k\varepsilon^p)^{6/(1-p)^2}\log^6 n)$.*

Proof. Combining the above two lemmas shows that the result of the algorithm has the desired accuracy. The reconstruction time can be broken down into the time to build J from the coefficients and the time to estimate the weight of each j in J. Building J takes time $O(k''\log n)$, since it requires a linear pass over the results of the measurements. To choose the location to find estimates quickly, we can build a vector $y = T_2\chi_J^T$ in time $O(|J|(k''\log n)^2)$, by selecting and summing the necessary columns. Then for each $j \in J$, we find some i such that $y_i = T_{2j,i} = 1$ and return the measurement $(T_2\Psi\mathbf{A})$ as $\hat{\theta}_j$. This takes $O((k''\log n)^2)$ time per coefficient. Lastly, picking the k largest of the estimated coefficients can be done with a linear pass over them. The dominating cost is $O(|J|(k''\log n)^2) = O((k''\log n)^3) = O((k\varepsilon^p)^{6/(1-p)^2}\log^6 n)$. □

The number of measurements is polynomial in $k, \log n$ (recall that p is fixed independent of n and $\mathbf{A}$). We have not optimized the various polynomial factors, but still, our methods will not yield less than k^4 measurements, due to the use of two collections of k-strongly separating sets. It is an open problem to further improve the number of measurements in explicit non-adaptive constructions. Note although we need to use p to define the measurements, we do not need the exact value of p. Rather, we need an upper bound on the true value of p (recall, the smaller the value of p, the faster the coefficients must reduce) — this is because our construction will simply take more coefficients than is necessary to get the required approximation accuracy.

2.2 Exponential Decay

As in the p-compressible case we state our results relative to the worst case error in the class for given α and C_α. In the case that $|\theta_i| = C_\alpha 2^{-\alpha i}$, we write $\|\mathbf{C}_{\text{opt}}^k\|_2^2 = \sum_{i=k+1}^n \theta_i^2$ as the worst case error over the class.

Measurements. The set of measurements we make is similar to the p-compressible case at the high level, but differs in the details. We set $k' = k + O(\frac{\log((k\log n)/\varepsilon)}{\alpha})$, and $k'' = O((k'\log n)^2)$ As before, we build $\mathcal{S}$, a k'-strongly separating collection of sets, and write T_3 as the concatenation of χ_{s_i} for all $S_i \in \mathcal{S}$ (k'' is chosen as the number of sets in the collection). However, we set $\mathcal{Q}$ to be a k''-separating collection of sets (not strongly separating), and write T_4 as its characteristic matrix. We form $T = (T_3 \otimes H) \bigcup T_4$.

Reconstruction algorithm. We recover a representation from the measurements from $T_3 \otimes H$ as before, to build a set J of identifiers. To make our estimates, we proceed iteratively to build $\hat{\theta}$, the vector of approximate coefficients. Initially $\hat{\theta} = 0$, and $M = \emptyset$. Let $j_1 \in (J \backslash M)$ satisfy $(J \backslash M) \cap Q_i = \{j_1\}$ (there will be at least one such Q_i and j_1). We set $\hat{\theta}_j = \chi_{Q_i}(\Psi\mathbf{A} - \hat{\theta})$ and $M = M \cup \{j_1\}$. We now proceed to find a new $j_2 \in (J \backslash M)$ with $(J \backslash M) \cap Q_{i'} = \{j_2\}$ as the next coefficient to estimate, and proceed until $J = M$. We then return the k highest estimated coefficients as before.

Lemma 4. *Let K' denote the set of the k' largest coefficients.*

1. $\forall j \in K' : \theta_j^2 \geq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2 \Rightarrow j \in J$
2. $\forall j \in J : |\hat{\theta}_j - \theta_j| \leq \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2$.

Proof. To show (1), we must bound the tail sums of coefficients of α-exponentially decaying signals. One can easily show that $\sum_{i=k+1}^n \theta_i^2 \leq c_\alpha 2^{-2\alpha k}$ and $(\sum_{i=k'+1}^n |\theta_i|)^2 = c'_\alpha 2^{-\alpha k'}$. Over the class of α-exponentially decaying signals, $(\sum_{i=k'}^n |\theta_i|)^2 \leq C_\alpha 2^{-\alpha(k'-k)}\|\mathbf{C}_{\text{opt}}^k\|_2^2$. Setting $k' = k + O(\frac{1}{\alpha}\log\frac{k}{\varepsilon})$ gives $(\sum_{i=k'}^n |\theta_i|)^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{C}_{\text{opt}}^k\|_2^2$. The remainder of the proof of (1) continues as in Lemma 2 (1).

To show (2), we scale ε by a factor of $O(k'')$. Note that this does not affect the asymptotic size of k'. This now ensures that the first coefficient j_1 is estimated with error $|\hat{\theta}_{j_1} - \theta_{j_1}| \leq \sum_{l=k'+1}^n |\theta_i| \leq \frac{\varepsilon}{k^{5/2}}\|\mathbf{C}_{\text{opt}}^k\|_2$. Now consider the estimation of the next coefficient j_2: it is possible that j_2 and j_1 occur in the same set $Q_{i'}$, in which case the error is bounded by $|\hat{\theta}_{j_2} - \theta_{j_2}| \leq |(\sum_{l\neq j_2, l\in Q_{i_2}} \theta_l) - \hat{\theta}_j| \leq \sum_{l\neq j_1, l\neq j_2, l\in Q_{i_2}} |\theta_l| + |\hat{\theta}_{j_1} - \theta_{j_1}| \leq \frac{2\varepsilon}{(k'\log n)^{5/2}}\|\mathbf{C}_{\text{opt}}^k\|_2$; else the error is bounded by $\frac{\varepsilon}{(k'\log n)^{5/2}}\|\mathbf{C}_{\text{opt}}^k\|_2$ as before. One can therefore show inductively that $|\hat{\theta}_{j_m} - \theta_{j_m}| \leq \frac{m\varepsilon}{5(k'\log n)^{5/2}}\|\mathbf{C}_{\text{opt}}^k\|_2$, and so, since $|J| \leq k'' = O((k'\log n)^2)$, we have $\forall j \in J. |\hat{\theta}_j - \theta_j| \leq \frac{\varepsilon}{5\sqrt{k}}\|\mathbf{C}_{\text{opt}}^k\|_2$, as required. □

Theorem 3. *We can construct a set of $O(k^2\,\text{polylog}(n))$ measurements in time polynomial in k and n. For any α-exponentially decaying signal $\mathbf{A}$, from these measurements of $\mathbf{A}$, we can return a representation $\mathbf{R}$ for $\mathbf{A}$ of at most k coefficients $\hat{\theta}$ under Ψ such that $\|\hat{\theta} - \theta\|_2^2 = \|\mathbf{R} - \mathbf{A}\|_2^2 < \|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \varepsilon\|\mathbf{C}_{\text{opt}}^k\|_2^2$. The time required to produce the coefficients from the measurements is $O(k^2\,\text{polylog}(n))$*

Proof. Using the results of Lemma 4 allows us to apply Lemma 3 and achieve the main theorem. For the time cost, we must first generate J, which takes time $O(k''\log n)$, and then iteratively build the estimates. This can be done efficiently in time $O(k''\,\text{polylog}(n))$ per coordinate, a constant number of operations on each of the $O(k''\,\text{polylog}(n))$ measurements.

For constant α and $\varepsilon = O(\text{poly}(1/n))$, we have $k' = O(k)$, $k'' = O((k \log n)^2)$ and the total number of measurements $= k'' \,\text{polylog}(n) = O(k^2 \log^{O(1)} n)$. □

k-sparse signals. For k-sparse signals, the previous method yields an explicit construction that guarantees exact reconstruction of the signal with $O(k^2 \log^{O(1)} n)$ measurements. The key observation is that after finding a superset J of the non-zero coefficients, the estimate of each of these coefficients is exact, since all coefficients not in J are zero. Thus,

Theorem 4. *We can construct a set of $O(k^2 \,\text{polylog}(n))$ measurements in time polynomial in k and n. For any k-sparse signal $\mathbf{A}$, from these measurements of $\mathbf{A}$, we can return a representation $\mathbf{R}$ for $\mathbf{A}$ of at most k coefficients $\hat{\theta}$ under Ψ such that $\|\hat{\theta} - \theta\|_2^2 = \|\mathbf{R} - \mathbf{A}\|_2^2 = \|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 = 0$. The time required to produce the coefficients from the measurements is $O(k^2 \,\text{polylog}(n))$*

3 Randomized Constructions

Here we focus on providing per-instance error estimates. For compressible signals (indeed this section works for arbitrary signals, but our focus is only on compressible signals as is standard in sparse approximation theory), one can give randomized constructions which guarantee to return a near-optimal representation for that signal, with high probability for each signal.

Transform Definition. Instead of using collections of sets with guaranteed separating properties, we make use of sets defined implicitly by hash functions to give a randomized separation property. We also use a random ± 1 valued vector to improve the accuracy of estimation of the coefficients. The necessary components are defined as follows:

Separation matrix M. M is a $0/1$ $s \times n$ matrix with the property that for every column, exactly one entry is 1, and the rest are zero. We will define M based on a randomly chosen function $g : [n] \to [s]$, where $\mathsf{Pr}[g(i) = j] = 1/s$ for $i \in [n], j \in [s]$. Hence, $M_{i,j} = 1 \iff g(i) = j$, and zero otherwise. The effect is to separate out the contributions of the coefficients: we say i is separated from a set K if $\forall j \in K . g(i) \neq g(j)$. For our proofs, we require that the mapping g is only three-wise independent, and we set $s = O(\frac{k \log n}{\varepsilon^2})$. This will ensure sufficient probability that any i is separated from the largest coefficients.

Estimation vector E. E is a ± 1 valued vector of dimension n so $\mathsf{Pr}[E_i = 1] = \mathsf{Pr}[E_i = -1] = \frac{1}{2}$. We will use the function $h : [n] \to \{-1, +1\}$ to refer to E, so that $E_i = h(i)$. For our proofs, we only require h to be four-wise independent.

Lastly, we compose T from M, H and E by: $T = M \otimes H \otimes E$.

Reconstruction Procedure. We consider each set of inner-products generated by the row M_j. When composed with $(H \otimes E)$, this leads to $1 + \log_2 n$ inner products, $x_0 \ldots x_{\log n} = (T\Psi\mathbf{A})_{j(1+\log n)} \ldots \theta'_{(j+1)(1+\log n)-1}$. From this, we attempt to recover a coefficient i by setting $i = \sum_{b=1}^{\log n} 2^{b-1} \frac{x_b^2 - \min\{x_b^2, (x_0 - x_b)^2\}}{\max\{x_b^2, (x_0 - x_b)^2\} - \min\{x_b^2, (x_0 - x_b)^2\}}$, and add i to our set of approximate coefficients,

$\hat{\theta}$. We estimate $\hat{\theta}_i = h(i)x_0$, and finally output as our approximate k largest coefficients those obtaining the k largest values of $|\hat{\theta}_i|$.

Lemma 5 (Coefficient recovery)**.** *(1) For every coefficient θ_i with $\theta_i^2 > \frac{\varepsilon^2}{25k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$, there is constant probability that the reconstruction procedure will return i (over the random choices of g and h).*
(2) We obtain an estimate of θ_i as $\hat{\theta}_i$ such that $(\theta_i - \hat{\theta}_i)^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ with constant probability.

Proof sketch. The outline of the proof is as follows: for each coefficient θ_i with $\theta_i^2 > \frac{\varepsilon^2}{25k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$, we show that there is constant probability that it is correctly recovered. Let

$$x_b = (\Psi'\mathbf{A})_{g(i)(1+\log n)+b} = \textstyle\sum_{g(j)=g(i)} H_{j,b}h(j)\theta_j.$$

One can show that $\mathsf{E}(x_b^2) \leq H_{i,b}\theta_i + O(\frac{\varepsilon^2}{k\log n})\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ and $\mathsf{Var}(x_b^2) \leq O(\frac{\varepsilon^2}{k\log n}\theta_i^2 H_{i,b}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 + \frac{\varepsilon^4}{k^2\log^2 n}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^4)$. Using the Chebyshev inequality on both x_b^2 and $(x_0 - x_b)^2$, and rearranging it can then be shown that $\mathsf{Pr}[\theta_i^2 - H_{i,b}(x_b^2) - (1 - H_{i,b})(x_0 - x_b)^2 \leq \frac{\theta_i^2}{2}] \leq \frac{2}{9\log n}$ and $\mathsf{Pr}[(1 - H_{i,b})x_b^2 + H_{i,b}(x_0 - x_b)^2 \geq \frac{\theta_i^2}{2}] \leq \frac{2}{9\log n}$. Combining these two results enables us to show that $\mathsf{Pr}[\frac{x_b^2 - \min\{x_b^2,(x_0-x_b)^2\}}{\max\{x_b^2,(x_0-x_b)^2\} - \min\{x_b^2,(x_0-x_b)^2\}} \neq H_{i,b}] \leq \frac{4}{9\log n}$. Thus, the probability that we recover i correctly is at least $\frac{5}{9}$.

For (2), we consider $\hat{\theta}_i = h(i)x_0 = h(i)\sum_{g(j)=g(i)} h(j)\theta_j$. One can easily verify that $\mathsf{E}(\hat{\theta}_i) = \theta_i$ and $\mathsf{Var}(\hat{\theta}_i) = \mathsf{E}(\sum_{g(j)=g(i),j\neq i}\theta_j^2)$. We argue that with constant probability none of the k largest coefficients collide with i under g, and so in expectation assuming this event $\mathsf{Var}(\hat{\theta}_i) = \frac{1}{s}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$. Applying the Chebyshev inequality to this, we obtain

$$\mathsf{Pr}[|\hat{\theta}_i - \theta_i| > \sqrt{\tfrac{\varepsilon^2}{9k}}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2] < \frac{\mathsf{Var}(\hat{\theta}_i)}{\frac{\varepsilon^2}{9k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2} \leq \frac{1}{9\log n}.$$

showing (2) with (better than) constant probability. □

Lemma 6 (Failure probability)**.** *By taking $O(\frac{ck\log^3 n}{\varepsilon^2})$ measurements we obtain a set of estimated coefficients $\hat{\theta}_i$ such that $(\theta_i - \hat{\theta}_i)^2 \leq \frac{\varepsilon^2}{25k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ with probability at least $1 - \frac{1}{n^c}$.*

Proof. In order to increase the probability of success from constant probability per coefficient to high probability over all coefficients, we will repeat the construction of T several times over using different randomly chosen functions g and h to generate the entries. We take $O(c\log n)$ repetitions: this guarantees that the probability of not returning any i with $\theta_i^2 > \frac{\varepsilon^2}{25k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ is n^{-c}, polynomially small. We also obtain $O(c\log n)$ estimates of θ_i from this procedure, one from each repetition of T. Each is within the desired bounds with constant probability at least $\frac{7}{8}$; taking the median of these estimates amplifies this to high probability using a standard Chernoff bounds argument. T has $m = s(\log n + 1) = O(\frac{k\log^2 n}{\varepsilon^2})$ rows, $O(c\log n)$ repetitions gives the stated bound. □

Theorem 5. *We can construct a dictionary $\Psi' = T\Psi$ of $O(\frac{ck\log^3 n}{\varepsilon^2})$ vectors, in time $O(cn^2\log n)$. For any signal $\mathbf{A}$, given the measurements $\Psi'\mathbf{A}$, we can find a representation $\mathbf{R}$ of $\mathbf{A}$ under*

Ψ such that with probability at least $1-\frac{1}{n^c}$ $\|\mathbf{R}-\mathbf{A}\|_2^2 \leq (1+\varepsilon)\|\mathbf{R}_{\text{opt}}^k-\mathbf{A}\|_2^2$. *The reconstruction process takes time* $O(\frac{c^2k\log^3 n}{\varepsilon^2})$.

The proof follows by combining the results of Lemma 5 with those of Lemma 3 to get the main result. We modify Lemma 3 to use $\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2$ in place of $\|\mathbf{C}_{\text{opt}}^k\|_2$; the proof is essentially the same. It is easy to verify that the number of coefficients identified by the first part of the reconstruction process is $O(\frac{ck\log^2 n}{\varepsilon})$ (taking time linear in m). We find an accurate estimate of each recovered coefficient by taking the median of $O(c\log n)$ estimates of each one.

If we spend linear time or more on reconstruction, we can work with fewer measurements.

Theorem 6. *We can construct a dictionary* $\Psi' = T\Psi$ *of* $O(\frac{ck\log n}{\varepsilon^2})$ *vectors, in time* $O(cn^2\log n)$. *For any signal* $\mathbf{A}$, *given the measurements* $\Psi'\mathbf{A}$, *we can find a representation* $\mathbf{R}$ *of* $\mathbf{A}$ *under* Ψ *such that with probability at least* $1-\frac{1}{n^c}$ $\|\mathbf{R}-\mathbf{A}\|_2^2 \leq (1+\varepsilon)\|\mathbf{R}_{\text{opt}}^k-\mathbf{A}\|_2^2$. *The reconstruction process takes time* $O(cn\log n)$.

Proof sketch. The construction here is similar to our main randomized result, but we do not use H and reduce s by a $\log n$ factor. Using only the separation and estimation matrices, we estimate *each* of the n coefficients, and take the k largest of them as before. By a similar argument to Lemma 5 (2), each coefficient is estimated with accuracy $\frac{\varepsilon^2}{25k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$, and we can again apply Lemma 3. □

Tolerance to Error. Several recent works have shown that compressed sensing-style techniques allow accurate reconstruction of the original signal even in the presence of error in the measurements (i.e. omission or distortion of certain θ_i's). We adopt the same model of error as [3, 22] and show:

Lemma 7. *1. If a fraction* $\rho = O(1)$ *of the measurements are chosen at random to be corrupted in an arbitrary fashion, we can still recover a representation* $\mathbf{R}$ *with error* $\|\mathbf{R} - \mathbf{A}\|_2^2 \leq (1+\varepsilon)\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ *in time* $O(cn\log n)$.
2. If only a $\rho = O(\log^{-1} n)$ *fraction of the measurements are corrupted we can recover a representation* $\mathbf{R}$ *with error* $\|\mathbf{R} - \mathbf{A}\|_2^2 \leq (1+\varepsilon)\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ *in time* $O(\frac{kc^2\log n}{\varepsilon^2})$.

Proof. 1. Consider the estimation of each coefficient in the process outlined in Theorem 6. Estimating θ_i takes the median of $O(\log n)$ estimates, each of which is accurate with constant probability. If the probability of an estimate being inaccurate or an error corrupting it is still constant, then the same Chernoff bounds argument guarantees accurate reconstruction. As long as ρ is less than a constant (say, 1/10) then every coefficient is recovered with error $\varepsilon\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2$, with high probability.

2. Consider the recovery of θ_i from T. We will be able to recover i provided the previous conditions hold, and additionally the some set of $\log n$ measurements of θ_i are not corrupted (we may still be able to recover i under corruption, but we pessimistically assume that this is not the case). Provided $\rho \leq 1/(3\log n)$ then each set of $\log n$ measurements are uncorrupted with constant probability at least 2/3 and hence we can recover i with high probability. The same argument for estimating θ_i accurately holds as in case (1). □

Note that the results in [22, 3] consider the exact recovery of a signal by taking $\Omega(n)$ measurements, and so do not compare to our result above of approximately recovering a signal using $o(n)$ measurements. Our randomized construction is also resilient to other models of error, such as the measurements being perturbed by some random vector of bounded weight (details omitted for brevity).

4 Concluding Remarks

We have presented a simple combinatorial approach of two sets of group tests with different separation properties that yields the first known polynomial time explicit construction of a non-adaptive transformation matrix and a reconstruction algorithm for the Compressed Sensing problem. Our approach yields other results including sublinear reconstruction, improved approximation in error and others. Given the excitement about Compressed Sensing in the Applied Mathematics community, we expect many new results soon. The main open problem is to reduce the number of measurements used by explicit algorithms: our result here gives a cost polynomial in k, which is not close to the linear factor k in the existential results of [4, 9, 22]. For the case of k-sparse signals, (which have no more than k nonzero coefficients) Indyk has very recently developed a set of measurements, linear in k in number (but has other superlogarithmic factors in n) [17].

Another outstanding question concerns the ability to choose the basis to reconstruct in after the measurements have been made. Due to the random structure of the measurement matrices in prior work [4, 9, 22], this "universality property" comes almost "for free": it just has to be incorporated into the linear program to be solved. In contrast, our reconstruction techniques rely on the fact that the measurements are sparse vectors in the desired basis, and so they do not naturally have the universality property. It remains open to extend our results to allow the basis to be specified after measurements have been made.

Acknowledgments. We thank Ron Devore, Ingrid Daubechies, Anna Gilbert and Martin Strauss for explaining compressed sensing.

References

[1] A. Akavia, S. Goldwasser, and S. Safra. Proving hard-core predicates by list decoding. In *IEEE Conference on Foundations of Computer Science*, pages 146–157, 2003.

[2] E. Candès, J. Romberg, and T. Tao. Stable signals recovery from incomplete and inaccurate measurements. `http://www.acm.caltech.edu/~emmanuel/papers/StableRecovery.pdf`, 2005. Unpublished Manuscript.

[3] E. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. In *IEEE Conference on Foundations of Computer Science*, 2005.

[4] E. Candès and T. Tao. Near optimal signal recovery from random projections and universal encoding strategies. Technical Report math.CA/0410542, arXiv, `http://arxiv.org/abs/math.CA/0410542`, 2004.

[5] A. Clementi, A. Monti, and R. Silvestri. Selective families, superimposed codes, and broadcasting on unknown radio networks. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[6] G. Cormode and S. Muthukrishnan. What's hot and what's not: Tracking most frequent items dynamically. In *Proceedings of ACM Principles of Database Systems*, pages 296–306, 2003.

[7] G. Cormode and S. Muthukrishnan. Towards an algorithmic theory of compressed sensing. Technical Report 2005-25, Center for Discrete Mathematics and Computer Science (DIMACS), 2005.

[8] R. Devore and G. G. Lorentz. *Constructive Approximation*, volume 303. Springer Grundlehren, 1993.

[9] D. Donoho. Compressed sensing. `http://www-stat.stanford.edu/~donoho/Reports/2004/CompressedSensing091604.pdf`, 2004. Unpublished Manuscript.

[10] D-Z Du and F.K. Hwang. *Combinatorial Group Testing and Its Applications*, volume 3 of *Series on Applied Mathematics*. World Scientific, 1993.

[11] A. Gilbert, S. Guha, P. Indyk, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Fast, small-space algorithms for approximate histogram maintenance. In *Proceedings of the ACM Symposiun on Theory of Computing*, pages 389–398, 2002.

[12] A. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss. Near-optimal sparse Fourier representation via sampling. In *Proceedings of the ACM Symposiun on Theory of Computing*, 2002.

[13] A. Gilbert, S. Muthukrishnan, and M. Strauss. Improved time bounds for near-optimal sparse Fourier representations. In *SPIE Conference on Wavelets*, 2005.

[14] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. `http://homepages.cae.wisc.edu/~jhaupt/Publications/RandomProjections05.pdf`, 2005. Unpublished Manuscript.

[15] IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. `http://www.icassp2005.com/`.

[16] P. Indyk. Explicit constructions of selectors and related combinatorial structures, with applications. In *Proceedings of ACM-SIAM Symposium on Discrete Algorithms*, 2002.

[17] P. Indyk. Personal communication, 2005.

[18] Integration of Sensing and Processing, 2005. http://www.ima.umn.edu/2005-2006/W12.5-9.05/.

[19] W.H. Kautz and R.R. Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on on Information Theory*, 10:363–377, 1964.

[20] E. Kushilevitz and Y. Mansour. Learning decision trees using the fourier spectrum. *SIAM Journal on Computing*, 22(6):1331–1348, 1993.

[21] Compressed sensing website. http://www.dsp.ece.rice.edu/CS/.

[22] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. http://www.math.ucdavis.edu/~vershynin/papers/ecc.pdf, 2005. Unpublished Manuscript.

[23] SPARS05 Signal Processing with Adaptative Sparse Structured Representations, 2005. http://www.irisa.fr/manifestations/2005/SPARS05/.

[24] Y. Tsaig and D. Donoho. Extensions of compressed sensing. http://www-stat.stanford.edu/~donoho/Reports/2004/ExtCS-10-22-04.pdf, 2004. Unpublished Manuscript.