DIMACS Technical Report 2005-25
September 2005

Towards an Algorithmic Theory of 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 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 Ψ for $\mathbb{R}^n$. Recently, there has been tremendous excitement about the novel direction of *Compressed Sensing* [10] where the reconstruction can be done with very few—$\tilde{O}(k)$—linear measurements over a modified dictionary Ψ' if the information of the signal is concentrated in k coefficients over an orthonormal basis Ψ. These results have reconstruction error on any given signal that is optimal with respect to a broad class of signals. In a series of papers and meetings over the past year, a theory of Compressed Sensing has been developed by mathematicians.

We develop an algorithmic perspective for the Compressed Sensing problem, showing that Compressed Sensing results resonate with prior work in Group Testing, Learning theory and Streaming algorithms. Our main contributions are new algorithms that present the most general results for Compressed Sensing with $1 + \epsilon$ approximation on *every* signal, faster algorithms for the reconstruction, as well as succinct transformations of Ψ to Ψ'.

1 Introduction

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 length n and $\psi_i \perp \psi_j$.[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] By Parseval's equality, $\sum_i \theta_i^2 = \langle \mathbf{A}, \mathbf{A} \rangle = \|\mathbf{A}\|_2^2$, i.e. the "energy" (sum of squares of values) of the signal is preserved under transformation by an orthonormal basis. In the area of sparse approximation theory [9], one seeks representations of $\mathbf{A}$ that are sparse, ie., 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 Parseval's equality, this is equivalently $\|\theta(\mathbf{A}) - \theta(\mathbf{R})\|_2^2$. The optimal k representation of $\mathbf{A}$ under Ψ, $\mathbf{R}^k_{\text{opt}}$ takes k coefficients with the largest $|\theta_i|$'s. From now on (for convenience of reference only), we reorder the vectors in the dictionary so $|\theta_1| \geq |\theta_2| \geq \ldots \geq |\theta_n|$. The error is $\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}\|_2^2 = \sum_{i=k+1}^n \theta_i^2$. Study of sparse approximation problems under different dictionaries is a mature area of Mathematics. There are three interesting cases:

- *k-support case.* If the signal has at most k non-zero coefficients[3] under Ψ, $\mathbf{R}^k_{\text{opt}}$ will have zero error ($\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 = 0$) and hence, $\mathbf{A}$ can be exactly reconstructed.
- *p-Compressible case.* In the area of sparse approximation theory, one typically studies functions that are compressible wrt Ψ. Specifically the coefficients have a power-law decay: for some $p \in (0, 1)$, $|\theta_i| = O(i^{-1/p})$ for a constant C. Consequently, $\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}\|_2^2 \leq C' k^{1-2/p}$ for some constant $C' = C'(C, p)$.
- *General case.* If $\mathbf{A}$ is arbitrary, for a given k, $\mathbf{R}^k_{\text{opt}}$ will have some arbitrary error.

Recently, there has been excitement in the area of Sparse Approximation Theory regarding what is called *Compressed Sensing*. The classical theory above says we should seek all linear measurements $\langle \mathbf{A}, \psi_i \rangle$'s so that the signal can be reconstructed. But if most of the "information" in $\mathbf{A}$ is concentrated in only a small number k of these, could one take fewer linear measurements? A priori we do not know which of the i's will be of our interest without looking at the signal. The recent excitement arises from several independently discovered results that show that Ψ can be transformed into a smaller dictionary Ψ' of $O(k \log n)$ vectors in $\mathbb{R}^n$ so that $\|\mathbf{A} - \mathbf{R}\|_2^2 \leq C' k^{1-2/p}$, for p-compressible signals. The error is within constant factors of the optimal k-term representation over the class of all such compressible signals. Donoho [10] called this Compressed Sensing since the size of the dictionary Ψ' (number of vectors in Ψ') is significantly smaller than n; it is also near-optimal in terms of the size of Ψ' since one needs at least k linear measurements to measure even the k non-zero coefficients in a signal with k-support. In a series of papers by three groups of researchers—Donoho,

[1] Examples of such basis are *standard* where $\psi_{i,j} = \delta_{i,j}$ where $\delta_{i,j}$ is the Kronecker delta function; *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, all these bases can be thought of as the standard basis.

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

[3] That is, for some $K \subset [n], |K| = k$, $i \notin K \Rightarrow \theta_i = 0$. In this case, we say θ has *support* k.

Candes and Tao, Rudelson and Vershynin—together with others, the theory of compressed sensing has been developed during the past year. The words they use to describe the results are "surprising", "remarkable", etc. The website [17] tracks the rapidly growing number of followup papers, both experimental and mathematical. These results have numerous applications to signal processing, error correction, and other fundamental areas, which adds to the tremendous excitement that is brewing [19].

We visit the problem of compressed sensing as algorithmicists. From an algorithmicist's point of view, one may wonder how the results above in Sparse Approximation Theory are reflected in prior work in Algorithms. There are three concrete directions that provide potential precursors. Finding the k non-zero coefficients among n in the k-support case reminds one of "group testing" to find k defective items from n identical-looking items that has been extensively studied in Discrete Mathematics and Algorithms [11] and is intimately related to Combinatorial Designs. Finding k representations optimally wrt an orthonormal basis such as the discrete Fourier basis reminds one of sampling for Fourier transform estimation in Learning and Complexity theory [16, 1, 14]. Finally, finding k representations using few linear measurements reminds one of work on data stream algorithms that use few inner products and small space for finding k-piecewise constant histograms [13] and finding top k elements [6, 7]. Still, it is difficult to pin down exactly the relationship between these results from different communities because there are a number of different aspects to these problems, and each community has its own concerns and style. What is needed is a more systematic formulation of the problems and results.

Further, viewed as an algorithmicist, the nature of results in Compressed Sensing seem unusual. One would seek a representation $\mathbf{R}(\mathbf{A})$ for the signal $\mathbf{A}$ that is provably accurate for the $\mathbf{A}$ (eg., $\|\mathbf{A} - \mathbf{R}^k\|_2^2 \leq f\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}(\mathbf{A})\|_2^2$ for some factor f), rather than having accuracy relative to the optimal over the class of all signals with some property (eg., $\|\mathbf{A} - \mathbf{R}^k\|_2^2 \leq C'k^{1-2/p}$) since for a given signal $\mathbf{A}$, $\|\mathbf{A} - \mathbf{R}^k_{\text{opt}}(\mathbf{A})\|_2^2$ may be much smaller than $C'k^{1-2/p}$. Finally, as algorithmicists, one quantitatively cares about a systematic study of the resources—space and randomness used and running time—needed for Compressed Sensing.

In this paper, we attempt to formalize the different *algorithmic* aspects of the Compressed Sensing problem. As it turns out, the problem is related to all of the three directions—group testing, Fourier transform estimation and data stream algorithms—we mentioned above and partial results may be obtained from using prior work. In addition, our main contributions are new algorithms for compressed sensing. They are the most general algorithmic results known for the compressed sensing problem, since they focus on obtaining error bounds for the instance of the signal, rather than a whole class.

2 The Compressed Sensing Problem

The compressed sensing problem can be thought of as having three parts.

1. *Dictionary Transform.* From the orthonormal basis Ψ, build a set Ψ' (m vectors of length n).
2. *Encoding.* Vectors $\mathbf{A}$ are "encoded" by Ψ', to give a set of coefficients $\theta'_i = \langle \psi'_i, \mathbf{A} \rangle$.
3. *Decoding.* Given the m values θ'_i, recover a representation of $\mathbf{A}$ under Ψ.

In evaluating the quality of such a scheme, we can focus on the following attributes:

- *Size of* Ψ': This goes to the heart of the Compressed Sensing problem: given a desired accuracy, how many measurements are required to give an accurate reconstruction of $\mathbf{A}$. This is lower bounded by the information needed to do the reconstruction even if we do not consider the computational requirements of doing the transformation or decoding. At least k measurements are necessary to compute the best k term representation, and even for k-support signals, a $k\log(n/k)$ lower bound follows from group testing. Consequently, one asks how close can we come to these bounds?

- *Error guarantee*: What is the error guarantee: is it with respect to the optimal for the given instance of the signal (called *instance-optimal*) or is it optimal wrt to the worst case error over a class of functions (class-optimal) such as the error $C'k^{1-2/p}$ for the p-compressible class of functions. From an algorithmicist's point of view, one prefers instance-optimal solutions since any given signal could have a best representation with much smaller error than the worst case over its entire class. For the k-support case, these notions coincide since the class-optimal error is zero.

- *Reconstruction cost*: how much time is needed for reconstruction? This is in particular critical in applications. For example, there is no requirement in the specification of the problem that the output be a sparse representation of $\mathbf{A}$ (i.e. with bounded support). This is because we only seek the representation to be close to optimal in error and there is no prescribed requirement on the sparseness of the output. Thus, decoding may even take $\Omega(n)$ time. However, since the signal has sparse representation, it is desirable to be efficient and have decoding only depend on the size (number of vectors) of Ψ'.

- *Failure model*: With what probability does the construction fail to give the required accuracy? Does the success depend on the signal, or will Ψ' work for *all* possible signals? Although all constructions of Ψ' are probabilistic, if the failure probability is exponentially small in n and k then, for a sufficiently restricted class of signals, there must exist Ψ' that works for any signal in the class. Such "non-adaptive" transformations are desirable and have led to some wonderment in the Compressed Sensing field. Otherwise, the failure probability is typically polynomially small (i.e. n^{-c} for some constant c).

- *Transformation cost*: What are resource bounds for the transformation? How much time is needed for the transformation and how succinctly can Ψ' be described. "Succinctness" depends on how many bits are needed to write down Ψ' when Ψ is the standard basis. Notice that one does not need $O(mn)$ bits since Ψ' can be implicitly specified, e.g. using hash functions. The minimum number of bits needed is $\log(mn)$.

Fundamentally, most algorithms proposed so far for Compressed Sensing are identical: they rely on the Basis Pursuit (BP) method, which is simply to solve the linear program to find the vector $\mathbf{A}'$ minimizing $\|\Psi'\mathbf{A}' - \Psi'\mathbf{A}\|_1$. Further, Ψ' is typically chosen as a random matrix, whose entries are indepent and identically distributed (iid) as Gaussian, Bernoulli or ± 1 [21]. Where the results vary is in the analysis showing the failure model and the size of Ψ' needed. In contrast, our main results are as follows:

Ref.	Signal	Error	Size of Ψ'	Reconstruct Time	Succinctness	Failure model
[5]	k-sparse	class	$O(k \log n)$	$\Omega(nk)$ LP solve	$O(kn \log n)$	probabilistic
[20]	k-sparse	class	$O(k \log n)$	$\Omega(k^2)$	$O(kn \log n)$	probabilistic
[18]	k-sparse	class	$O(k \log n/k)$	$\Omega(nk)$ LP solve	$O(kn \log n/k)$	non-adaptive
Here	k-sparse	class	$O(k \log^2 n)$	$O(k \log^2 n)$	$O(\log n)$	probabilistic
Here	k-sparse	class	$O(k^2 \log n \log n/k)$	$O(k^2 \log n \log n/k)$	$O(k \log n/k)$	non-adaptive
[4]	p-compressible	class	$O(k \log n)$	$\Omega(nk)$ LP solve	$O(kn \log n)$	probabilistic
[10]	p-compressible	class	$O(k \log n)$	$\Omega(nk)$ LP solve	$O(kn \log n)$	non-adaptive
[18]	p-compressible	class	$O(k \log n/k)$	$\Omega(nk)$ LP solve	$O(kn \log n/k)$	non-adaptive
Here	p-compressible	instance	$O(k^{\frac{3-p}{1-p}} \log^2 n)$	$O(k^{\frac{4-2p}{1-p}} \log^3 n)$	$O(k^{\frac{2-p}{1-p}} \log n)$	non-adaptive
[13]	general	instance	$\Omega(\frac{1}{\epsilon^3} k \log n)$	$\Omega(\frac{1}{\epsilon^3} k^2 \log n)$	$O(\log^2 n)$	probabilistic
Here	general	instance	$O(\frac{k}{\epsilon} \log^{5/2} n)$	$O(\frac{k}{\epsilon} \log^{5/2} n)$	$O(\log^2 n)$	probabilistic

Table 1: Comparison of prior and current work. "LP Solve" denotes the time requires to solve a linear program on $\Omega(nk)$ variables.

- For arbitrary signals, we show a new randomized construction that produces a Ψ' with $O(\frac{k \log^{5/2} n}{\epsilon})$ vectors and present an algorithm that recovers a representation $\mathbf{R}^k$ with k non-zero coefficients of vectors from Ψ so that $\|\mathbf{R}^k - \mathbf{A}\|_2^2 \leq (1+\epsilon)\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ with high probability. Our results are thus instance optimal. Reconstruction is linear in the number of measurements (up to constant factors), and the cost depends only logarithmically on n.

- For p-compressible and k-sparse signals, we go further and show the existence of a single transformation that will work for all inputs in Section 4. This matches Donoho's results in that the dictionary is nonadaptive, but our construction is also instance-optimal and hence requires more measurements. Further, it is easy to verify that the generated matrix has the nonadaptive property. These results are naturally resilient to error: provided at most a small fraction of measurements are corrupted, we can still recover a representation of the signal to the same error bounds as before. This complements the most recent compressed sensing results [3], which show similar results for the Basis Pursuit methods.

Our results improve on prior results in compressed sensing in that they work for non-compressible (general) signals: they are instance optimal, in contrast to all prior compressed sensing results, which are class optimal over the restricted classes. But our results are weaker since they require more inner products than the $O(k \log n/k)$ obtained in prior work for compressible signals. We significantly improve results in running times since previous works mostly rely on solving a linear program of size $\tilde{O}(nk)$ [4] which in general takes time cubic in the program size. Even for the k support case (only k non-zero coefficients) the previous results rely on Orthogonal Matching Pursuit [20] (which is at least quadratic in k from preliminary empirical analysis) and the explicit group testing construction in [12] takes

[4]The notation $\tilde{O}(\cdot)$ and $\tilde{\Omega}(\cdot)$ supresses factors of $\log n$ when these are small in comparison to other factors

time $\Omega(k^4)$. Another result [8] is similar, but does not provide any bounds on the error. Lastly, we show that we need only polylogarithmic bits of randomness to construct Ψ' (i.e. it has a very succint representation) so construction of Ψ' is fast.

The Fourier transform estimation results [14, 15] can be applied to the Compressed Sensing Problem to get a result similar to above, but our result has much improved bounds in terms of $k, \epsilon^{-1}, \log n$, and works for any orthonormal dictionary. Similarly, the histogram/wavelet data stream algorithm in [13] can be applied to get a similar result, but the dependency on parameters is polynomially higher than here: the dependency on ϵ, for example, is at least cubic. Here, we improve the running time by polynomial factors in k, ϵ, and our algorithm avoids the "greedy pursuit" method required to decode and identify the coefficients: a single pass over the measurements suffices to reconstruct. Similarly, some frequent element algorithms on data streams [6, 7, 8] can get similar results by combining estimation techniques there with the lemmas we prove here. But still, our results here give the stronger bounds on the accuracy of recovered coefficient values, and much reduced decoding time than the $\Omega(n)$ cost of directly applying these methods.

Our results are obtained by matrix transformations that are at the heart of group testing algorithms, Fourier sampling methods and data stream algorithms. But besides extracting all the ingredients needed for out result here, our main result is achieved by an improved analysis of these ingredients. For example, we estimate coefficients where the prior methods show accuracy depending on $\|\mathbf{A}\|$, but we prove it in terms of the norm of the error $\|\mathbf{A}-\mathbf{R}^k_{\text{opt}}\|$ which is significantly tighter and the basis for all our results here.

Comparison of existing methods and the results we present here on each of the main attributes is given in Table 1. We consider which of the three models of signals is assumed, and whether the error is class- or instance-optimal. We then report the number of measurements (i.e. size of Ψ'), the reconstrucion time (which is either the time to solve a linear program which is at least linear in n, or sublinear in n), and the succinctness of Ψ' in bits.

3 Our Algorithms

The goal is to produce a set of m (row) vectors Ψ'. 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, ie, $\Psi' = T\Psi$, 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.

3.1 Defining the Transform Matrix

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. We will define a transform

T whose goal is to recover k coefficients approximately so that the error from using these k coefficients is within a $(1+\epsilon)$ factor of using the k largest coefficients from θ. Later, we use standard repetition techniques to boost this to arbitrarily high probability for *all* large coefficients.

We expose the key conditions we use to be able to find and estimate the necessary coefficients. These are *macroseparation*: that the coefficients get spread out from one another so we can identify one without "interference" from others; *microseparation*: in order to identify a coefficient, we use a standard Hamming code-like structure to find its identity, akin to non-adaptive group testing [11]; and *estimation*: to get a good estimate of the coefficient, we need to show an accurate estimator for this quantity. We build the transformation matrix T from the linear combination of three pieces, each of which achieves one of the above goals.

- **Macroseparation matrix S.** S 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 S 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, $S_{i,j} = 1 \iff g(i) = j$, 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 $s = \frac{16k \log^{1/2} n}{\epsilon}$. This will ensure sufficient probability that any i is separated from the largest coefficients.

- **Microseparation matrix H.** H is the $0/1$ $(1 + 2\lceil \log_2 n)\rceil \times n$ matrix derived from the $(1 + \lceil \log_2 n)\rceil \times 2^{\lceil \log_2 n \rceil}$ Hamming code matrix by taking the first n columns[5]. Let M denote the Hamming matrix, we now set $H_{2i} = M_i$ and $H_{2i-1} = H_0 - M_i$ (recalling that H_0 is a row of all ones). Define $bit(i, j)$ as the function that returns the jth bit of i in the binary representation of i. Formally, $H_{0,j} = 1, H_{2i,j} = bit(i, j)$, and $H_{2i-1,j} = 1 - bit(i, j)$ for $i = 1 \ldots \log n$.

- **Estimation vector E.** E is a ± 1 valued vector of length 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.

To build T from S, H and E, we combine them with tensor product-like linear operator $\otimes$.

Definition 1. *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 $(V \otimes W)_{iv+l,j} = V_{i,j} W_{l,j}$. We compose T from S, H and E by: $T = S \otimes H \otimes E$.*

We will let $m = s(2 \log n + 1)$ and observe that T is an $m \times n$ matrix.

Lemma 1 (Encoding cost and succinctness)**.** *T can be specified in $O(\log n)$ bits. The set of m vectors $\Psi' = T\Psi$ can be constructed in time $O(n^2 \log n)$.*

[5]From now on, we will drop base and ceiling notation, and let $\log n$ stand for $\lceil \log_2 n \rceil$.

Proof. Observe that T is fully specified by g and h. Since both of these hash functions are four-wise independent, they can be specified with $O(\log n)$ bits. Naïvely, one could construct Ψ' by constructing T explicitly, and then multiplying T by Ψ to get Ψ'. Using the best known matrix multiplication algorithms, this would take time superquadratic in n and $\log n$. However, we can use the sparsity and structure of T to allow us to construct Ψ' faster. Consider each entry of Ψ in turn. This contributes to $O(\log n)$ entries in Ψ': S ensures that each entry is picked once, and then $(H \otimes E)$ splits this into $O(\log n)$ inner products. Given $\Psi_{i,j}$, we know this can contribute to $\Psi'_{l,j}$ for l in the range $(2\log n + 1)g(i) \ldots (2\log n + 1)g(i) + 2\log n$, and for each of these entries we add one of $\{-\Psi_{i,j}, 0, +\Psi_{i,j}\}$, depending on the values of $h(j)$ and H_i. Hence, by taking a linear pass over Ψ, we can construct $\Psi' = T\Psi$, using $O(\log n)$ time for each entry in Ψ. This gives the overall running time.

For more structured dictionaries Ψ which are themselves sparse (e.g. Haar wavelet basis), or highly structured (e.g. Discrete Fourier basis) one could improve this running time further, nearly linear in the dimensions of Ψ' and hence almost optimally; we omit full details. □

3.2 Reconstruction of Coefficients

Decoding Procedure. We consider each set of inner-products generated by the row S_j. When composed with $(H \otimes E)$, this leads to $1+2\log_2 n$ inner products, $(T\Psi\mathbf{A})_{j(1+2\log n)} \ldots \theta'_{(j+1)(1+2\log n)-1}$ which we denote $x_0 \ldots x_{2\log n}$. From this, we attempt to decode a coefficient by comparing x_{2l}^2 with x_{2l-1}^2: if $x_{2l}^2 > x_{2l-1}^2$ then we set $b_l = 1$; else we set $b_l = 0$. We then find $i = \sum_{l=1}^{\log n} b_l 2^{l-1}$, 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 2 (Coefficient recovery)**.** *For every coefficient θ_i with $\theta_i^2 > \frac{\epsilon}{2k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$, there is constant probability that the decoding procedure will return i.*

Proof. Consider some $\theta_i = (\Psi\mathbf{A})_i$ satisfying the stated condition. Each test between x_{2l}^2 and x_{2l-1}^2 is an attempt to probe the lth bit of the binary represenation of i: if $bit(i,l)$ is 1, then x_{2l} is a linear combination of θ_i and some other coefficients, while x_{2l-1} is a linear combination of other coefficients. If the contributions of other coefficients are sufficiently small in comparison to θ_i, then x_{2l}^2 will be larger than x_{2l-1}^2, and so we will correctly recover the lth bit of i. If $bit(i,l) = 0$, then the argument is symmetric. We repeat this procedure for all $\log n$ values of l to recover the full binary description of i. Our argument is that the separation matrix causes the contribution of other coefficients to x_{2l} and x_{2l-1} is small; in particular, we argue that there is constant probability that none of the k largest coefficients contribute to these values.

In order to formalize this intuition, we consider the contribution of θ_i to the results of the inner products. Let x denote the vector of results of the inner-products involving θ_i, i.e. $x_l = \theta'_{g(i)(1+2\log n)+l}$. Formally,

$$x_l = (\Psi'\mathbf{A})_{g(i)(1+2\log n)+l} = ((S \otimes H \otimes E)\theta)_{g(i)(1+2\log n)+l} = (S_{g(i)} \otimes H_l \otimes E)\theta = \sum_{g(j)=g(i)} H_{j,l}h(j)\theta_j$$

Consider x_{2l} and x_{2l-1}, and assume that $bit(i,l) = 1$ (the other case is symmetric). We consider the probability that $x_{2l}^2 < x_{2l-1}^2$. That is, the probability that we fail to recover the

lth bit of i correctly. First, we define a random variable $X = x_{2l}^2 - \theta_i^2$ and analyse expectation and variance of this variable.

$$\begin{aligned}
\mathsf{E}(X) &= \mathsf{E}(\textstyle\sum_{g(j)=g(i)} H_{j,2l} h(j)\theta_j)^2) \\
&= \mathsf{E}(\textstyle\sum_{g(j)=g(i),j\neq i} H_{j,2l} h^2(j)\theta_j^2 + \sum_{g(j)=g(q)=g(i),j\neq q\neq i} H_{j,2l}H_{q,2l} h(j)h(q)\theta_j\theta_q) \\
&= \mathsf{E}(\textstyle\sum_{g(j)=g(i),j\neq i} H_{j,l}\theta_j^2) = \frac{1}{s}\sum_{j\neq i} H_{j,2l}\theta_j^2 = \frac{\epsilon}{16k\log^{1/2} n}\sum_{j\neq i} H_{j,2l}\theta_j^2 \\
\mathsf{Var}(X) &= \mathsf{E}(\textstyle\sum_{g(j)=g(i)} H_{j,2l} h(j)\theta_j)^4) - \mathsf{E}(X)^2 \\
&= \mathsf{E}(\textstyle\sum_{g(j)=g(q)=g(i)} 6H_{j,2l}H_{q,2l} h^2(j)h^2(q)\theta_j^2\theta_q^2) - \mathsf{E}(\textstyle\sum_{g(j)=g(i)} H_{j,2l}\theta_j^2)^2 \\
&\leq \mathsf{E}(\textstyle\sum_{g(j)=g(q)=g(i),j\neq q\neq i} 4\theta_j^2\theta_q^2)
\end{aligned}$$

This relies on the independence properties of g and h so that terms with odd powers of h are zero in expectation (eg for $j \neq q$, $\mathsf{E}(h(j)h(q)) = 0$), and that g distributes values uniformly. Further, we argue that, with at least constant probability, none of the k largest coefficients are mapped to the same value as i under g (unless θ_i itself is one of the k largest coefficients). By the pairwise independence of g, $\mathsf{Pr}[g(i) = g(j), i \neq j] = \frac{1}{s} = \frac{\epsilon}{16k\log^{1/2} n}$ and so the probability of this event not happening for $1 \leq j \neq i \leq k$ is $1 - \frac{\epsilon}{16\log^{1/2} n}$ by the union bound. For reasonable values of $\epsilon \leq 1$ and $\log n \geq 4$, this is at least $\frac{31}{32}$. Under the assumption that this condition holds, we can write the expectation and variance of X in terms of the optimal error $\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ (under direct application of prior works in small space representations, the bound would instead be in terms of $\|\mathbf{A}\|_2^2$).

$$\mathsf{E}(X) = \frac{\epsilon\sum_{j>k,j\neq i} H_{j,2l}\theta_j^2}{16k\log^{1/2} n} \leq \frac{\epsilon\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2}{16k\log^{1/2} n} \text{ and } \mathsf{Var}(X) \leq \textstyle\sum_{g(j)=g(q),j,q\geq k} 4\theta_j^2\theta_q^2 \leq \frac{\epsilon^2\cdot 2\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^4}{16^2k^2\log n}.$$

Applying the Chebyshev inequality, we obtain

$$\mathsf{Pr}\left[|(x_{2l}^2 - \theta_i^2) - \mathsf{E}(X)| > \frac{3\epsilon}{16k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2\right] \leq \frac{\frac{\epsilon^2}{128k^2\log n}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^4}{(\frac{3\epsilon}{16k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2)^2} \leq \frac{2}{9\log n}.$$

Rearranging, we have $\mathsf{Pr}[|\theta_i^2 - x_{2l}^2| > \frac{\epsilon}{4k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] \leq \mathsf{Pr}[x_{2l}^2 < \theta_i^2 - \frac{\epsilon}{4k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] \leq \frac{2}{9\log n}$. By a virtually identical analysis, one can compute that $\mathsf{Pr}[x_{2l-1}^2 > \frac{\epsilon}{4k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] \leq \frac{2}{9\log n}$. Combining these two results, and using the fact that $\theta_i^2 > \frac{\epsilon}{2k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$, we have that

$$\begin{aligned}
\mathsf{Pr}[x_{2l}^2 < x_{2l-1}^2] &= \mathsf{Pr}[x_{2l}^2 - x_{2l-1}^2 < 0] < \mathsf{Pr}[x_{2l}^2 - x_{2l-1}^2 < \theta_i^2 - \frac{\epsilon}{2k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] \\
&\leq \mathsf{Pr}[x_{2l}^2 < \theta_i^2 - \frac{\epsilon}{4k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] + \mathsf{Pr}[x_{2l-1}^2 > \frac{\epsilon}{4k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2] \leq \frac{4}{9\log n}
\end{aligned}$$

Hence, testing whether $x_{2l}^2 > x_{2l-1}^2$ in order to discover which contains θ_i gives the correct answer with probability at least $1 - \frac{4}{9\log n}$. Combining the results of $\log n$ tests to get the full identity i therefore succeeds with at least constant probability ($\frac{5}{9}$), using the union bound. The total probability of success is constant, since the probability of failing is at most $\frac{1}{32} + \frac{4}{9} = \frac{137}{288}$ for each i (there is a collision with one of the k largest coefficients, or one of the comparisons gives the wrong answer). □

Lemma 3 (Accurate estimation)**.** *We obtain an estimate of θ_i as $\hat{\theta}_i$ such that $|\theta_i^2 - \hat{\theta}_i^2| \leq \frac{\epsilon}{2k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ and $(\theta_i - \hat{\theta}_i)^2 \leq \frac{\epsilon}{k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ with constant probability.*

Proof. Once we have identified i, we must additionally return an estimate of θ_i with the correct bounds. We now show that we can estimate θ_i from the results of the inner products so that $|\theta_i - \hat{\theta}_i| \leq \sqrt{\frac{\epsilon}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2}$. For this we need only the macroseparation and estimation properties of T. We return to the vector x of values that contain a contribution from θ_i, and consider $x_0 = \sum_{g(j)=g(i)} h(j)\theta_j$. Recall that we set $\hat{\theta}_i = h(i)x_0$. One can easily verify that $\mathsf{E}(\hat{\theta}_i) = \theta_i$ and $\mathsf{Var}(\hat{\theta}_i) = \mathsf{E}(\sum_{g(j)=g(i), j\neq i} \theta_j^2)$. Again, we argue that with constant probability none of the k largest coefficients collide with i under g, and so in expectation assuming this event $\mathsf{Var}(\hat{\theta}_i) = \frac{\epsilon}{16k\log^{1/2} n}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$. Applying the Chebyshev inequality to this, we obtain

$$\mathsf{Pr}[|\hat{\theta}_i - \theta_i| > \sqrt{\tfrac{\epsilon}{k}}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2] < \frac{\mathsf{Var}(\hat{\theta}_i)}{\frac{\epsilon}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2} = \frac{1}{16\log^{1/2} n}.$$

For the second accuracy bound, observe that if $\hat{\theta}_i = h(i)x_0$ then $\hat{\theta}_i^2 = x_0^2$. We have already shown that $\mathsf{Pr}[|x_{2l}^2 - \theta_i^2| > \frac{\epsilon}{2k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2] \leq \frac{2}{9\log n}$. This also holds for x_0 using the same proof. Provided $\log n \geq 4$, both properties hold simultaneously with probability at least $1 - (\frac{1}{32} + \frac{1}{18}) = \frac{263}{288}$. □

Lemma 4 (Failure probability)**.** *By taking $O(\frac{ck\log^{5/2} n}{\epsilon})$ measurements we obtain an estimate of θ_i as $\hat{\theta}_i$ for every coefficient $1 \leq i \leq n$, such that $|\theta_i^2 - \hat{\theta}_i^2| \leq \frac{\epsilon}{2k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ and $(\theta_i - \hat{\theta}_i)^2 \leq \frac{\epsilon}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ with probability at least $1 - n^{-c}$.*

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 $\theta_i^2 > \frac{\epsilon}{2k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ 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. Lastly, note that if $\theta_i^2 < \frac{\epsilon}{2k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ then (notionally) setting $\hat{\theta}_i = 0$ satisfies both conditions on the estimate $\hat{\theta}_i$, so the accuracy bounds hold for all coefficients $1 \leq i \leq n$. T has $m = s(\log n + 1) = O(\frac{k\log^{3/2} n}{\epsilon})$ rows, $O(c\log n)$ repetitions gives the stated bound. □

Lemma 5 (Reconstruction accuracy)**.** *Given $\hat{\theta}(\mathbf{A}) = \{\hat{\theta}_i(\mathbf{A})\}$ such that both $|\hat{\theta}_i^2 - \theta_i^2| \leq \frac{\epsilon}{2k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ and $(\hat{\theta}_i - \theta_i)^2 \leq \frac{\epsilon}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ for all i, picking the k largest coefficients from $\hat{\theta}(\mathbf{A})$ gives a $(1+\epsilon)$ approximation of the optimal k term representation of $\mathbf{A}$.*

Proof. As stated in the introduction, the error from picking the k largest coefficients exactly is $\|\theta(\mathbf{A}) - \theta(\mathbf{R}^k_{\text{opt}})\|_2^2 = \sum_{i=k+1}^n \theta_i^2$ (where we index the θ_is in decreasing order of size). We will write $\hat{\phi}_i$ for the i largest approximate coefficient, and ϕ_i for its exact value. Let $\pi(i)$ denote the mapping such that $\phi_i = \theta_{\pi(i)}$. Picking the k largest approximate coefficients has energy error

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

Consider i such that $i > k$ but $\pi(i) < k$: this corresponds to a coefficient that belongs in the top k but whose estimate leads us to not choose it. The threshold for being included in the top-k coefficients is $\hat{\phi}_k^2$. So we have $\hat{\phi}_i^2 \leq \hat{\phi}_k^2 \leq \hat{\phi}_{\pi(i)}^2$, since $\hat{\phi}_{\pi(i)}^2$ corresponds to some item that is in the top-k approximate coefficients. But using our bounds on estimation and rearranging: $\phi_i^2 \leq \phi_{\pi(i)}^2 + \frac{\epsilon}{k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$. So the missing coefficient cannot be very much larger than one that is included.

$$\begin{aligned}
\|\mathbf{R} - \mathbf{A}\|_2^2 &\leq \sum_{i \leq k, \pi(i) \leq k} \frac{\epsilon}{k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \sum_{i > k, \pi(i) \leq k} (\theta_i^2 + \frac{\epsilon}{k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2) + \sum_{i > k, \pi(i) > k} \theta_i^2 \\
&\leq \sum_{\pi(i) \leq k} \frac{\epsilon}{k}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \sum_{i > k} \theta_i^2 \leq \epsilon\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 + \sum_{i > k} \theta_i^2 \\
&= (1 + \epsilon)\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2
\end{aligned}$$

□

Lemma 6 (Reconstruction cost)**.** *The decoding process takes time* $O(\frac{c^2 k \log^{5/2} n}{\epsilon})$.

The proof of the lemma is straightforward: we just need to generate the set of $|K| = O(\frac{ck \log^{3/2} n}{\epsilon})$ coefficients from running the first part of the decoding process (taking time linear in m), then find an accurate estimate of each decoded coefficient by taking the median of $O(c \log n)$ estimates of each. We note that one can trade off a $O(\log^{3/2} n)$ factor in the number of inner products taken if we allow the decoding time to increase to $O(cn \log n)$, by using the macro-separation and estimation matrices to iteratively estimate all n coefficients, and taking the k largest of them. □

Finally, we combine the preceding series of lemmas to get the main theorem:

Theorem 1. *With probability at least* $1 - n^{-c}$*, and in time* $O(c^2 \frac{k}{\epsilon} \log^{5/2} n)$ *we can find a representation* $\mathbf{R}$ *of* $\mathbf{A}$ *under* Ψ *such that* $\|\mathbf{R} - \mathbf{A}\|_2^2 \leq (1 + \epsilon)\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ *and* $\|\mathbf{R}\|$ *has support* k*. The dictionary* $\Psi' = T\Psi$ *has* $O(\frac{ck \log^{5/2} n}{\epsilon})$ *vectors, and is constructed in time* $O(cn^2 \log n)$*;* T *is represented with* $O(c^2 \log n)$ *bits.x*

Corollary 1. *If* $\mathbf{A}$ *has support* k *under* Ψ *then we can find a representation* $\mathbf{R}$ *under* Ψ *where the size of* Ψ' *is* $O(k \log^2 n)$ *so that with probability at least* $1 - n^{-c}$*, we find the exact representation* $\mathbf{R}$ *of* $\mathbf{A}$*. If* $\mathbf{A}$ *is* p*-compressible under* Ψ *then we can build* Ψ' *of size* $O(k \log^{5/2} n)$ *so that with probability at least* $1 - n^{-c}$*, we can find a representation* $\mathbf{R}$ *of* $\mathbf{A}$ *under* Ψ *such that* $\|\mathbf{R} - \mathbf{A}\|_2^2 \leq (1 + \epsilon)C'k^{1-2/p}$.

The corollary follows almost immediately from Theorem 1 by substituting the bounds for $\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$ for the two cases and setting $\epsilon = O(1)$. For the k-sparse case, we can shave the $\log^{1/2} n$ factor off the size of s, because there are no coefficients causing collisions beyond the k largest.

4 Non-adaptive Dictionary Transformation

p-Compressible case. In the p-compressible case the coefficients (sorted by magnitude and normalized) obey $|\theta_i| = O(i^{-1/p})$ for appropriate scaling constants and some parameter p. Previous work has focused on the cases $0 < p < 1$ [4, 10]. Integrating shows that $\sum_{i=k+1}^{n} \theta_i^2 = \|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2 = O(k^{1-2/p})$.

Because this model essentially states that distribution of coefficients has a tail of small weight, we can use this to create a matrix T which is good for *any* signal $\mathbf{A}$ obeying this property (our previous results were with high probability). 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 so small that it is still within our allowable error bounds.

Theorem 2. *There exists a set of $O(k^{\frac{3-p}{1-p}} \log^2 n)$ vectors Ψ' so that* any *p-compressible signal can be recovered with error $O(\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2)$. Moreover, there is a Las Vegas-style randomized algorithm to find such a set of inner products.*

Proof. We use the same construction of T and decoding procedure as before, but we change the parameters and take advantage of the information from the p-compressible case about the guaranteed decay in size of the coefficients to be able to guarantee accurate recovery of the original signal. Observe that the square of the (absolute) sums of coefficients after removing the top k' is $(\sum_{i=k'+1}^{n} |\theta_i|)^2 = O(k'^{2-2/p})$. We set this equal to $\theta_k^2 = O(k^{-2/p})$ and so $k' = O(k^{\frac{1}{(1-p)}})$ Having chosen k', we choose $s = \Omega(k')$ so that the probability of any of the top k' coefficients colliding with any θ_i for $i \in K \subset [n]$ (where $|K| = kk'$) under g is at most a small constant ($\frac{1}{8}$, say). We take enough repetitions of T so that in at least half the repetitions i is not separated from any of the top-k' coefficients with probability $o(n^{-kk'})$. This can be done using $O(\log n^{kk'})$ repetitions, using a standard Chernoff bounds argument. Now consider the number of ways of picking K and the top-k' coefficients: there are $O(n^{kk'})$ such possibilities. The probability of failure on any of these choices is $O(n^{-kk'} n^{kk'}) = O(1)$. Consequently, there must exist a set of repetitions of T with this property (moreover, drawing such a set randomly succeeds with at least constant probability). Hence, we can give a Las Vegas algorithm to find such a set: build one and test whether, for all choices of K and the top-k' coefficients that it has this "deterministic strong separation" property. If so, accept it, else repeat until one is found. (This is in contrast to previous work where it is unclear how to check whether a given transform matrix has the necessary non-adaptive properties).

We can now state a deterministic version of Lemma 2: given a set of inner products with the deterministic separation property, we can guarantee to recover the top-k coefficients. The observation is that for each of the top-k coefficients, there is now guaranteed to be a set of $(1 + 2\log n)$ inner products x (as in Lemma 2) such that none of the top-k' coefficients collide under g. Hence, the only way we could fail to recover a coefficient i is if $(\sum_{g(j)=g(i), j>k'} h(j)\theta_j)^2 > \theta_i^2$: if the items from the k' tail colliding with i under g are enough to give the wrong answer in comparison to θ_i^2. But this sum is at most $(\sum_{j=k'+1}^{n} |\theta_j|)^2$, which by our choice of k' is less than θ_k^2, so for $i \leq k$ this event cannot happen, no matter what values g and h take.

This ensures that we can recover the identity of the top k coefficients. However, in total we recover a set K of coefficients, with $|K| \leq kk'$, and we need to estimate the size of each of these coefficients accurately, using a deterministic version of Lemmas 3 and 4. The error in estimation each such coefficient can be bounded to $\frac{1}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2$ if in at least half of the estimates of θ_i we avoid the top k' coefficients, since k' satisfies $(\sum_{j=k'+1}^{n} |\theta_i|)^2 \leq \frac{1}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2 = O(k^{-2/p})$. Since the maximum error in estimation is bounded as $|\hat{\theta}_i^2 - \theta_i^2| \leq O(\frac{1}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2)$ and $(\hat{\theta}_i - \theta_i)^2 \leq O(\frac{1}{k}\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2)$, we can apply Lemma 5 in conjunction with the fact that we have the true top-k coefficients amongst our set K, and conclude that we recover a representation of $\mathbf{A}$ as $\mathbf{R}$ with error $O(\|\mathbf{R}^k_{\text{opt}} - \mathbf{A}\|_2^2)$.

The overall number of inner products needed is $O(kk'^2 \log^2 n)$: $O(k' \log n)$ to guarantee constant probability of separation for each repetition of T, and $O(kk' \log n)$ repetitions of T to give deterministic strong separation. Thus the overall number of inner products is $O(k^{1+\frac{2}{1-p}} \log^2 n) = O(k^{\frac{3-p}{1-p}} \log^2 n)$. □

k-support case. When θ has support k, we can apply the above analysis more tightly, since estimation can be done exactly.

Corollary 2. *There exists a set of $O(k^2 \log n \log n/k)$ vectors Ψ' so that* any *signal with support k can be recovered exactly, and there exists a Las Vegas-style randomized algorithm to find such a set in expected time $\tilde{O}(k^3 n^k)$.*

Proof sketch. Since there are only k non-zero coefficients, we only have to ensure that there is at least one repetition of k where each coefficient does not collide with any of the others. When this happens, we can decode i and θ_i exactly, since we can verify that $x_{2l} = \theta_i$ and $x_{2l-1} = 0$ or vice-versa for each l. If there is a collision, then there will be some value of l for which both x_{2l} and x_{2l-1} are non-zero. So we need $s = O(k)$ to ensure constant probability of separation in each repetition, H and E as before, and $O(k \log n/k)$ repetitions, for a total dictionary size of $O(k^2 \log^2 n)$. As before, not only does the exponentially small probability of failure ensure that there must exist a construction with the necessary separation property for all possible signals with support k, but we can generate and test randomly generated instances in order to find one. □

5 Concluding Remarks

There has been tremendous excitement in the Mathematics community for compressed sensing since its christening in [10]. We have taken an algorithmicist's view and proposed new algorithms for compressed sensing that are instance-optimal, have fast reconstruction and are quite simple. In Mathematics the momentum seems to indicate even more interest in Compressed Sensing in the future and application of those ideas to foundations (eg, error correction as in [3]) or to practice (as explored in experimental studies of [21, 2]). Our algorithmic results here will be of interest to them. A natural set of open problems is to find the tightest bounds for compressed sensing under the set of metrics we have formalized here (size of Ψ', reconstruction cost, error guarantee etc.).

One of the interesting lively directions in Compressed Sensing is working error-resiliently [3, 18]. 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, 18] in order to show that a certain level of error resilience comes "for free" with our construction.

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

Proof. Consider the recovery of θ_i from T. We will be able to recover i provided the previous conditions hold, and additionally the $\log n$ 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 all $\log n$ measurements are uncorrupted with constant probability at least $2/3$ and hence we can recover i with constant probability. Similarly, 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 this also holds with constant probability. Combining these, we are able to recover the signal to the same level of accuracy using $O(\frac{k\log^{5/2}}{\epsilon})$ measurements, if $\rho \leq 1/(3\log n)$. □

We can strengthen the bounds on ρ to $O(1)$, at the expense of higher decoding cost, by directly estimating all θ_i as suggested after Lemma 6.

The construction is also resilient to other models of error, such as the measurements being perturbed by some random vector of bounded weight. Provided the weight of the perturbation vector is at most $e\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$, for some constant e, say, then it is straightforward to modify the earlier proofs to tolerate such error, since in expectation one can argue that the error introduced into each measure is bounded by $\frac{(1+e)\epsilon}{k\log^{3/2} n}\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$, and so the overall accuracy will be $(1+(1+e)\epsilon)\|\mathbf{R}_{\text{opt}}^k - \mathbf{A}\|_2^2$.

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 and J. Romberg. Practical signal recovery from random projections. `http://www.acm.caltech.edu/~emmanuel/papers/PracticalRecovery.pdf`, 2005. Unpublished Manuscript.

[3] E. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming, 2005. In FOCS.

[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] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles and optimally sparse decompositions. http://www.acm.caltech.edu/~emmanuel/papers/OptimalRecovery.pdf, 2004. Unpublished Manuscript.

[6] M. Charikar, K. Chen, and M. Farach-Colton. Finding frequent items in data streams. In *Procedings of the International Colloquium on Automata, Languages and Programming (ICALP)*, pages 693–703, 2002.

[7] 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.

[8] G. Cormode and S. Muthukrishnan. What's new: Finding significant differences in network data streams. In *Proceedings of IEEE Infocom*, 2004.

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

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

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

[12] L. Gasiniec and S. Muthukrishnan. Explicit construction of small space group testing designs and applications, 2005. Unpublished Manuscript.

[13] 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.

[14] 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.

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

[16] Y. Mansour. Randomized interpoloation and approximation of sparse polynomials. *SIAM Journal of Computing*, 24(2), 1995.

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

[18] 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.

[19] CSCAMM workshop on sparse representation in redundant systems, 2005.

[20] J. Tropp and A. Gilbert. Signal recovery from partial information via orthogonal matching pursuit. http://www-personal.umich.edu/~jtropp/papers/TG05-Signal-Recovery.pdf, 2005. Unpublished Manuscript.

[21] 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.