Sparse Approximation and Compressed Sensing Using the Reed-Muller Sieve

Robert Calderbank, Stephen Howard, Sina Jafarpour, and Jeremy Kent

***Abstract*—This paper introduces the Witness-Averaging Algorithm for sparse reconstruction using the Reed Muller sieve. The Reed-Muller sieve is a deterministic measurement matrix for compressed sensing. The columns of this matrix are obtained by exponentiating codewords in the quaternary second order Reed Muller code of length N. For $k = \tilde{O}\left(N\right)$, the Witness-Averaging improves upon prior methods for identifying the *support* of a k-sparse vector by removing the requirement that the signal entries be independent, and by providing computational efficiency. It also enables local detection; that is, the proposed algorithm detects the presence or absence of a signal at any given position in the data domain without explicitly reconstructing the entire signal. Reconstruction is shown to be resilient to noise in both the measurement and data domains; the *average-case* ℓ_2/ℓ_2 error bounds derived in this paper are tighter than the *worst-case* ℓ_2/ℓ_1 bounds arising from random ensembles.**

***Index Terms*—Compressed Sensing, Reed-Muller Sieve, Support-Localized Detection, Delsarte-Goethals Codes, The Probabilistic Method**

I. Introduction

The central goal of compressed sensing is to capture attributes of a signal using very few measurements. In most work to date, this broader objective is exemplified by the important special case in which the measurement data constitute a vector $f = \Phi\alpha + e$, where Φ is an $N \times \mathcal{C}$ matrix called the *sensing matrix*, α is a vector in $\mathbb{C}^{\mathcal{C}}$ which can be well-approximated by a vector with at most k non-zero entries (a k-sparse vector), and e is additive measurement noise.

The role of random measurement in compressive sensing (see [1] and [2]) can be viewed as analogous to the role of random coding in Shannon theory. Both provide worst-case performance guarantees in the context of an adversarial signal/error model. In the standard paradigm, the measurement matrix is required to act as a near isometry on all k-sparse signals (this is the Restricted Isometry Property or RIP introduced in [3]). Although it is known that certain probabilistic processes generate $N \times \mathcal{C}$ measurement matrices that satisfy the RIP with high probability, there is no practical algorithm for verifying whether a given measurement matrix has this property. Storing the entries of a random sensing matrix may also require significant resources.

Basis Pursuit [1], [4], [5], Matching Pursuit [6]–[9], and Belief Propagation algorithms [10]–[12] can be used to recover a k-sparse signal from the N measurements using a RIP matrix. However, these algorithms only provide sparse approximation and there is no guarantee that they recover the support of the original sparse signal in the presence of noise.

The Reed Muller sieve is a deterministic sensing matrix. The columns are obtained by exponentiating codewords in the quaternary second order Reed Muller code; they are uniformly and very precisely distributed over the surface of an N-dimensional sphere. Coherence between columns reduces to properties of these algebraic codes and we use these properties to show that recovery of k-sparse signals is possible with high probability.

The deterministic structure of the Reed-Muller sieve provides restrictive and efficient reconstruction algorithms. When the sparsity level $k = \tilde{O}\left(\sqrt{N}\right)$[1], recovery is possible using the "quadratic reconstruction algorithm" presented in [13] and the reconstruction complexity is only $\tilde{O}(k\,N)$. The prospect of designing matrices for which very fast recovery algorithms are possible is one of the motivations for deterministic compressive sensing. When the sparsity level $k = \tilde{O}(N)$ recovery is possible using the algorithm described in this paper.

Nevertheless, reconstruction of a signal from sensor data is often not the ultimate goal and it is of considerable interest in imaging to be able to deduce attributes of the signal from the measurements without explicitly reconstructing the full signal. In particular, there are many important applications where the objective is to identify the signal model (the support of the signal α). These include network anomaly detection where the objective is to characterize anomalous flows and cognitive

RC is with the Department of Computer Science, Duke University, Durham, NC 27707, email: *calderbk@cs.duke.edu*, SH is with the Defense Science and Technology Organisation, PO Box 1500, Edinburgh, Australia, email: *show@ee.unimelb.edu.au*, SJ is with the Department of Computer Science, Princeton University, Princeton NJ 08540, email: *sina@cs.princeton.edu*, JK is with the Department of Mathematics, Princeton University, Princeton NJ 08540, email: *jakent@princeton.edu*.

[1] Throughout this paper the notation $\tilde{O}$ is used to avoid rewriting the constant and poly-log terms.

radio where the objective is to characterize spectral occupancy. The Reed Muller sieve improves on results obtained by Candès and Plan [14] in that for $k = \tilde{O}(N)$ it is able to identify the signal model without requiring that the signal entries have independent random sign. We also show that the Reed Muller sieve is able to detect the presence or absence of a signal at any given position in the data domain without needing to first reconstruct the entire signal. The complexity of such detection is $N^2 \log N$. This makes it possible to quickly calculate thumbnail images and to zoom in on areas of interest.

There are two models for evaluating noise resilience in compressive sensing. We provide an average case error analysis for both the stochastic model where noise in the data and measurement domains is usually taken to be iid white Gaussian, and the deterministic model where the goal is to approximate a compressible signal. It is the geometry of the sieve, more precisely the careful design of coherence between columns of the measurement matrix, which provides resilience to noise in both the measurement and the data domain. Our analysis points to the importance of both the average and the worst-case coherence.

We show that the ℓ_2 error in reconstruction is bounded above by the ℓ_2 error of the best k-term approximation. This type of ℓ_2/ℓ_2 bound is tighter than the ℓ_2/ℓ_1 bounds arising from random ensembles [1], [7] and the ℓ_1/ℓ_1 bounds arising from expander-based ensembles [15], [16]. We emphasize that our error bound is for average-case analysis, whereas the results obtained by Cohen et. al. [17] show that worst-case ℓ_2/ℓ_2 approximation is not achievable unless $N = O(\mathcal{C})$.

Note that the average-case ℓ_2/ℓ_2 bound is also achievable by sensing matrices constructed from hash functions [18], [19]; however, the construction of those matrices is randomized, whereas the sensing matrices proposed in this paper have explicit constructions. Table I summarizes the comparison of the proposed Witness-Averaging Algorithm with prior work.

Outline. The rest of the paper is organized as follows: Section §II clarifies the notations used in the paper. In Section §III we introduce the Delsarte-Goethals sieves; we further analyze the average and worst-case coherence of these matrices. Section §IV tightens the StRIP bounds introduces by Calderbank, Howard and Jafarpour [13], and generalizes the results to arbitrary functions. In Section §V we introduce the witness-averaging algorithm and its sparse reconstruction guarantees. Sections §VI proves the main results of this paper. Section §VII provides experimental results comparing the proposed algorithm with the state of the art compressive sensing algorithms in the literature. Section §VIII concludes the paper.

II. Background and Notation

This Section introduces notation and reviews the theory of sparse reconstruction. In this paper we focus on average case analysis.

A. Notation

Given a vector $v = (v_1, \cdots, v_n)$ in $\mathbb{R}^n$, $\|v\|_2$ denotes the Euclidean norm of v, and $\|v\|_1$ denotes the ℓ_1 norm of v defined as $\|v\|_1 \doteq \sum_{i=1}^n |v_i|$. We further define $\|v\|_\infty \doteq \max\{|v_1|, \cdots, |v_n|\}$, and $\|v\|_{\min} \doteq \min\{|v_1|, \cdots, |v_n|\}$. Also the Hamming weight of v is defined as $\|v\|_0 \doteq \{i : v_i \neq 0\}$. Whenever it is clear from the context, we drop the subscript from the ℓ_2 norm. Also $v_{i\to j}$ denotes the vector v restricted to entries $i, i+1, \cdots, j$, that is $v_{i\to j} \doteq (v_i, v_{i+1}, \cdots, v_j)$. Let A be a matrix with rank r. We denote the conjugate transpose of A by $A^\dagger$. Let $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_r]$ denote the vector of the singular values of A. The spectral norm $\|A\|$ of a matrix A is the largest singular value of A: that is $\|A\| \doteq \|\boldsymbol{\sigma}\|_\infty$.

Throughout this paper we shall use the notation φ_j for the j^{th} column of the sensing matrix Φ; its entries will be denoted by $\varphi_j(x)$, with the row label x varying from 0 to $N-1$. In other words, $\varphi_j(x)$ is the entry of Φ in row x and column j. We denote the set $\{1, \cdots .\mathcal{C}\}$ by $[\mathcal{C}]$. Given a subset S of $[\mathcal{C}]$, the matrix obtained by restricting Φ to the columns in S is denoted Φ_S. The indicator function $\delta_{a,b}$ is defined by

$$\delta_{a,b} = \begin{cases} 1 & \text{if a=b} \\ 0 & \text{otherwise} \end{cases}$$

A vector $\alpha \in \mathbb{C}^{\mathcal{C}}$ is k-sparse if it has at most k non-zero entries. The support of the k-sparse vector α, denoted by Supp(α), contains the indices of the non-zero entries of α. Let $\pi = \{\pi_1, \cdots, \pi_{\mathcal{C}}\}$ be a uniformly random permutation of $[\mathcal{C}]$. Up to section §V we always assume that α is a k-sparse signal with Supp$(\alpha) = \{\pi_1, \cdots, \pi_k\}$. We further assume that conditioned on the support, the values of the k non-zero entries of α are sampled from a distribution which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^k$. The *Minimum to Average Ratio* (MAR) of a k-sparse signal is defined as $\mathrm{MAR}(\alpha) \doteq \frac{k\,\|\alpha\|^2_{\min}}{\|\alpha\|^2}$. This means that $\mathrm{MAR}(\alpha)$ is the ratio of the energy in the smallest nonzero entry of alpha to the average signal energy per nonzero entry.

Remark 1. *From Section §V onwards we consider general vectors $\alpha \in \mathbb{C}^{\mathcal{C}}$. We generalize the average case analysis to a non-sparse vector α as follows: Let $\pi = \{\pi_1, \cdots, \pi_{\mathcal{C}}\}$ be a uniformly random permutation of $[\mathcal{C}]$. Without loss of generality, we can assume that the entries of α are sorted by there magnitudes and the columns of the sensing matrix Φ are randomly permuted by π. Therefore, for every index i,*

TABLE I

Comparison of the Witness-Averaging Algorithm with prior approaches used in sparse recovery and compressed sensing. For each approach, only one reference is provided and the reader is encouraged to see that reference for further related work. To preserve space, all constants are dropped from the "Number of Measurements" and "Recovery Time" columns. A reconstruction algorithm provides ℓ_p/ℓ_q noise-tolerance if for every signal in the signal model, the ℓ_p error in recovery is bounded above by the ℓ_q error of the best k-term approximation of the signal.

Sensing Matrix	Recovery Algorithm	Number of Measurements	Recovery Time	Noise Tolerance	Explicit Construction	Signal Model
RS Codes [20]	Algebraic	k	k^2	No	Yes	Worst-Case
Gaussian [1]	Basis Pursuit	$k\log\left(\frac{\mathcal{C}}{k}\right)$	$\mathcal{C}^3$	ℓ_2/ℓ_1	No	Worst-Case
Gaussian [21]	Dantzig Selector	$k\log\left(\frac{\mathcal{C}}{k}\right)$	$\mathcal{C}^3$	ℓ_2/ℓ_1	No	Worst-Case
Gaussian [7]	Greedy	$k\log\left(\frac{\mathcal{C}}{k}\right)$	$k\mathcal{C}\log\left(\frac{\mathcal{C}}{k}\right)$	ℓ_2/ℓ_1	No	Worst-Case
Expander [15]	Greedy	$k\log\left(\frac{\mathcal{C}}{k}\right)$	$\mathcal{C}\log\left(\frac{\mathcal{C}}{k}\right)$	ℓ_1/ℓ_1	No	Worst-Case
RM Frame [22]	LASSO	$k\log(\mathcal{C})$	$\mathcal{C}^3$	ℓ_2/ℓ_2	Yes	Average-Case
Hashing [18]	Group Testing	$k\log^5(\mathcal{C})$	$k\log^5(\mathcal{C})$	ℓ_2/ℓ_2	No	Average-Case
Expander [15]	Greedy	$k\log^{\Omega(1)}\left(\frac{\mathcal{C}}{k}\right)$	$\mathcal{C}\log\left(\frac{\mathcal{C}}{k}\right)$	ℓ_1/ℓ_1	Yes	Worst-Case
Extractor [23]	Basis Pursuit	$k^{1.9}$	$\mathcal{C}^3$	ℓ_2/ℓ_1	Yes	Worst-Case
Toeplitz [24]	Basis Pursuit	k^2	$\mathcal{C}^3$	ℓ_2/ℓ_1	Yes	Worst-Case
RM sieve	Witness-Averaging	$k\log^2\mathcal{C}$	$k\mathcal{C}\log^2\mathcal{C}$	ℓ_2/ℓ_2	Yes	Average-Case

α_i (which is the i^{th} largest entry of α), corresponds to the π_i^{th} column of the matrix. The vector $\alpha_{1\to k}$ denotes the best k-term approximation of α.

Big O notation. Throughout the paper the notation $\preceq$ and $\succeq$ will be used to provide an upper bound on the growth rate of functions. Thus $A \preceq B$ and $B \succeq A$ if $A = O(B)$. We shall also use $\mathrm{Poly}(\mathcal{C})$ to denote $\mathcal{C}^{O(1)}$. More precisely, the term "with probability $1 - O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, $A \preceq B$" is equivalent to the statement "for every positive τ there exists a constant $\kappa(\tau)$ such that $\Pr\left[A > \kappa(\tau)B\right] \le \frac{1}{\mathcal{C}^\tau}$."

Group Theory. In this paper, we are interested in deterministic sensing matrices for which the columns form a group $\mathcal{G}$ under pointwise multiplication. The multiplicative identity is the column $\mathbf{1}$ with every entry equal to 1. The following property is fundamental.

Lemma 2. *If every row contains some entry not equal to* 1, *then the column group $\mathcal{G}$ satisfies $\sum_{g\in\mathcal{G}} g = 0$*

Proof: Given a row x and an element $f(x) \neq 1$, we have

$$f(x)\left(\sum_g g(x)\right) = \sum_g \left(f(x)g(x)\right) = \sum_g g(x).$$

■

B. Incoherent Dictionaries

An $N \times \mathcal{C}$ matrix Φ with normalized columns is called a dictionary. The two fundamental measures of coherence between the columns of Φ are defined as [25]:

- **Worst-case Coherence**:

$$\mu \doteq \max_{i\neq j}\left|\varphi_i^\dagger \varphi_j\right|.$$

- **Average Coherence**:

$$\nu \doteq \frac{1}{\mathcal{C}-1}\max_i\left|\sum_{j:j\neq i}\varphi_i^\dagger\varphi_j\right|.$$

The following results are due to Tropp [26] and show that with overwhelming probability the ℓ_0 minimization program successfully recovers the original k-sparse signal.

Theorem 3. *Assume the dictionary Φ satisfies $\mu \le \frac{c}{\log\mathcal{C}}$, where c is an absolute constant. Further assume $k \le \frac{c\,\mathcal{C}}{\|\Phi\|^2\log\mathcal{C}}$. Let S be a random subset of $[\mathcal{C}]$ of size k, and let Φ_S be the corresponding $N \times k$ submatrix. Then there exists an absolute constant c_0 such that*

$$\Pr\left[\left\|\Phi_S^\dagger\Phi_S - I\right\| \ge c_0\left(\mu\log\mathcal{C} + 2\sqrt{\frac{\|\Phi\|^2 k}{\mathcal{C}}}\right)\right] \le 2\,\mathcal{C}^{-1}.$$

Theorem 4. *Assume the dictionary Φ satisfies $\mu \le \frac{c}{\log\mathcal{C}}$, where c is an absolute constant. Further assume $k \le \frac{c\,\mathcal{C}}{\|\Phi\|^2\log\mathcal{C}}$. Let α be a k-sparse vector, such that the support of the k nonzero entries of α is selected uniformly at random. Then with probability $1 - O\left(\mathcal{C}^{-1}\right)$, α is the unique k-sparse vector mapped to $u = \Phi\alpha$ by the measurement matrix Φ.*

III. The Reed-Muller Sieve

A. Delsarte-Goethals Sieves

The Delsarte-Goethals sieve is a measurement matrix introduced by Calderbank, Howard, and Jafarpour [22], [27].

It is parametrized by two integers (m, r), where m is an odd number, r is between 0 and $\frac{m-1}{2}$. The matrix has $N \doteq 2^m - m - 1$ rows and $\mathcal{C} \doteq 2^{(r+1)m}$ columns. The rows of the sensing matrix Φ are indexed by the binary m-tuples x (where $x \notin \{0, 1, 2, 4, \cdots, 2^{m-1}\}$), and the $\mathcal{C}$ columns are indexed by matrices P, where P is an $m \times m$ binary symmetric matrix in the Delsarte-Goethals set $DG(m, r)$. The entry $\varphi_P(x)$ is given by

$$\varphi_P(x) = \frac{1}{\sqrt{N}} \imath^{xPx^\top}. \tag{1}$$

All entries in x and P are 0 or 1, but the exponent $xPx^\top$ is calculated in the ring of integers modulo 4. In fact the vector $(xPx^\top)$ is a codeword in the Delsarte-Goethals code (defined over the ring of integers modulo 4). The Delsarte-Goethals set $DG(m, r)$ is a binary vector space containing $2^{(r+1)m}$ binary symmetric matrices P with the property that the difference of any two distinct matrices has rank at least $m - 2r$. The first set $DG(m, 0)$ is the classical Kerdock set, and the last set $DG(m, {}^{(m-1)}/_{2})$ is the set of all binary symmetric matrices. We refer the interested reader to [28]–[30] and Chapter 15 of [31] for more information about binary symmetric matrices and subcodes of the second order Reed-Muller code.

Let $\hat{\Phi} = \sqrt{N}\Phi$ denote the unnormalized sensing matrix. We can divide the columns of $\hat{\Phi}$ into a set H indexed by the matrices in $DG(m, r)$ with zero diagonal, and a set D indexed by the matrices in the Kerdock set. The columns in H form a group under pointwise multiplication, and since we have excluded the rows that are a power of two, there exists no row in H with every entry equal 1. First, using the group property we calculate the average coherence of Φ.

Lemma 5. *The average coherence of a $DG(m, r)$ sieve is:*

$$\nu \doteq \max_i \left| \frac{1}{\mathcal{C}-1} \sum_{j:j\neq i} \varphi_i^\dagger \varphi_j \right| = \frac{1}{\mathcal{C}-1}.$$

Proof: Any column of $\hat{\Phi}$ can be written as a pointwise product hd with h in H and d in D. Average coherence with respect to hd is then

$$\frac{1}{N(\mathcal{C}-1)} \sum_{(h',d')\neq(h,d)} d^{-1}h^{-1}h'd'. \tag{2}$$

case 1: $d' \neq d$. In this case h' ranges over all elements of H, and using Proposition 2 we get

$$\frac{1}{N(\mathcal{C}-1)} \left[\sum_{d'\neq d} \mathbf{1}^\top d^{-1} \left(\sum_{h' \in \mathrm{H}} h^{-1}h' \right) d'\mathbf{1} \right]$$
$$= \frac{1}{N\left(\mathcal{C}-1\right)} \left[\sum_{d'\neq d)} \mathbf{1}^\top d^{-1} \left(\sum_{h'' \in \mathrm{H}} h'' \right) d'\mathbf{1} \right] = 0.$$

case 2: $d' = d$ and hence $h' \neq h$. In this case (2) reduces to

$$\frac{1}{N(\mathcal{C}-1)} \left[\mathbf{1}^\top d^{-1} \left(\sum_{h'\neq h} h^{-1}h' \right) d\mathbf{1} \right]. \tag{3}$$

Again, Proposition 2 implies that

$$\left(\sum_{h'\neq h} h^{-1}h' \right) + h^{-1}h = \sum_{h'\in \mathrm{H}} h^{-1}h' = \sum_{h''\in \mathrm{H}} h'' = 0.$$

Hence Equation 3 further reduces to

$$\frac{-1}{N\left(\mathcal{C}-1\right)} \left[\mathbf{1}^\top d^{-1} d\mathbf{1} \right] = \frac{-1}{\mathcal{C}-1}.$$

■

The next two results are generalizations of Proposition A.2 of [13] and are used in Section §VI to analyze the sparse reconstruction algorithm.

Lemma 6. *Let P be a binary symmetric $m \times m$ matrix and let E be the null space of P. Let X be an ℓ dimensional subspace of $\mathbb{F}_2^m$ and let $f \in \mathbb{F}_2^m$. If*

$$S = \sum_{x \in X+f} \imath^{xPx^\top + 2bx^\top} \text{ where } b \in \mathbb{F}_2^m,$$

then either $S = 0$ or

$$S^2 = \imath^{fPf^\top + 2bf^\top} \imath^{z_1Pz_1^\top + 2(b+fP)z_1^\top} 2^{\ell + dim(X\cap E)},$$

where $z_1 \in X$ is a solution to $z_1 P = d_P$ and d_P is the main diagonal of P.

Proof: For simplicity we first consider the case $f = 0$. We have

$$\begin{aligned} S^2 &= \sum_{x,y} \imath^{xPx^\top + yPy^\top + 2b(x+y)^\top} \\ &= \sum_{x,y} \imath^{(x+y)P(x+y)^\top + 2xPy^\top + 2b(x+y)^\top}. \end{aligned}$$

Changing variables to $z = x + y$ and y gives

$$\begin{aligned} S^2 &= \sum_z \imath^{zPz^\top + 2bz^\top} \sum_y (-1)^{(z+y)Py^\top} \\ &= \sum_z \imath^{zPz^\top + 2bz^\top} \sum_y (-1)^{(d_P + zP)y^\top}. \end{aligned}$$

Since the diagonal of P is contained in the row space of P, there exits a solution z_1 in $\mathbb{F}_2^m$ to $zP = d_P$. Note that if $e, f \in E$ then $ePe^\top + fPf^\top = (e+f)P(e+f)^\top \pmod 4$. If there is no solution z_1 in X to the equation $zP = d_P$, then $S = 0$. Otherwise

$$\begin{aligned} S^2 &= 2^\ell \sum_{e \in E\cap X} \imath^{(z_1+e)P(z_1+e)^\top + 2b(z_1+e)^\top} \\ &= 2^\ell \imath^{z_1Pz_1^\top + 2z_1b^\top} \sum_{e\in E\cap X} \imath^{ePe^\top + 2be^\top}. \end{aligned}$$

The map $e \rightarrow ePe^\top$ is a linear map from E to $2\,\mathbb{Z}_4$, so the numerator $ePe^\top + 2be^\top$ also determines a linear map. If this map is the zero map then $S^2 = 2^{\ell+\dim(E\cap X)}\imath^{z_1Pz_1^\top+2z_1b^\top}$, and if not then $S = 0$. The general case reduces to the case $f = 0$ since

$$\sum_{x\in X}\imath^{(f+x)P(f+x)^\top+2b(f+x)^\top} = \imath^{fPf^\top+2bf^\top}\sum_{x\in X}\imath^{xPx^\top+2(b+fP)x^\top}.$$

■

Lemma 7. *Let P, Q be two binary symmetric $m\times m$ matrices and $\mathcal{N}_Q$ and $\mathcal{N}_{P-Q}$ be the null spaces of Q and $P-Q$. If*

$$S = \sum_{a\in\mathbb{F}_2^m}\sum_{x\in\mathbb{F}_2^m}\imath^{aPa^\top+xQx^\top+2aQx^\top},$$

then

$$\left|S^2\right| \le 2^{2m}|\mathcal{N}_Q|\;|\mathcal{N}_{P-Q}|.$$

Proof: We have

$$S^2 = \sum_{\substack{a,b\\x,y}}\imath^{aPa^\top+bPb^\top+xQx^\top+yQy^\top+2aQx^\top+2bQy^\top}.$$

Changing variables to $z \doteq x+y$, y, $c \doteq a+b$, and b, implies that $S^2 =$

$$\sum_{\substack{b,c\\y,z}}\imath^{cPc^\top+2(b+c)Pb^\top+zQz^\top+2(y+z)Qy^\top+2(b+c)Q(y+z)^\top+2bQy^\top} \tag{4}$$

$$= \sum_{c,z}\imath^{cPc^\top+zQz^\top+2cQz^\top}\mathcal{T}(c,z),$$

with

$$\mathcal{T}(c,z) \doteq \sum_b(-1)^{(cP+d_P+zQ)b^\top}\sum_y(-1)^{(zQ+d_Q+cQ)y^\top}.$$

The term in (4) vanishes unless $cP + d_P + zQ = 0$ and $zQ + d_Q + cQ = 0$ simultaneously. Hence, we can rewrite Equation (4) as

$$S^2 = 2^{2m}\sum_{\substack{c,z\\(c+z)Q=d_Q\\c(P+Q)=d_P+d_Q}}\imath^{(c+z)Q(c+z)^\top+c(P-Q)c^\top}.$$

Write $c = c_1 + e$ with $c_1(P-Q) = d_P + d_Q$, $e(P-Q) = 0$, and $c+z = (c_2+z_2)+f$ with $(c_2+z_2)Q = d_Q$, $fQ = 0$. Then it follows from the triangle inequality that $\left|S^2\right|$ is at most

$$2^{2m}\left|\sum_f\imath^{fQf^\top}\right|\;\left|\sum_e\imath^{e(P-Q)e^\top}\right| = 2^{2m}|\mathcal{N}_Q|\;|\mathcal{N}_{P-Q}|.$$

■

Now we bound the worst-case coherence between the columns of Φ.

Lemma 8. *Let Φ be a $DG(m,r)$ sieve. Then*

$$\mu \doteq \max_{i\neq j}\left|\varphi_i^\dagger\varphi_j\right| \le \frac{1}{N^{\frac{1}{2}-\frac{r+1}{m}}}.$$

Proof: Lemma 6 proves that for every P and Q in $DG(m,r)$ set

$$\left|\sum_{x\in\mathbb{F}_2^m}\frac{1}{N}\imath^{x(P-Q)x^\top}\right| \le \frac{1}{N^{\frac{1}{2}-\frac{r}{m}}}.$$

Recall that we have excluded $m+1$ rows from the matrix indexed by $x = 0, 1, \cdots, 2^{m-1}$. However, every entry of the sensing matrix has magnitude $\frac{1}{\sqrt{N}}$. Hence, it follows from the triangle inequality that

$$\mu \le \frac{1}{N^{\frac{1}{2}-\frac{r}{m}}} + \frac{m+1}{N} < \frac{1}{N^{\frac{1}{2}-\frac{r+1}{m}}}.$$

■

Remark 9. *An argument similar to the one used in Lemma 8 can be used to generalize the upperbounds proved in Lemmas 6 and 7 to the case where the $m+1$ duplicate rows are removed from the sensing matrix.*

Remark 10. *In this paper, we have excluded the rows of the DG sieve which are a power of two. However, this subsampling can be removed if we add a phase shift $i^{\mathrm{wt}(\mathrm{d_P})}$ to the columns of the matrix. Here $\mathrm{wt}(\mathrm{d_P})$ denotes the Hamming weight of the diagonal of matrix P. Numerical experiments suggest that the phase-shifted Reed-Muller sieve has almost the same performance as the Reed-Muller sieve without phase shifts [32]. However, for simplicity, throughout this paper we only analyze the Reed-Muller sieve without phase shift, whose entries are given by Equation* (1).

B. Noise Shaping

1) Stochastic Noise Model: We have verified that for $m \le 17$ every $DG(m,r)$ sieve with $r \ge 2$ is a tight frame with redundancy $\frac{\mathcal{C}}{N}$ (see [22]). Note that when $m = 17$ the measurement matrix has $131,072$ rows. We conjecture that all such sieves are tight frames and we will analyze the statistical noise model under this assumption[2]. Therefore $\Phi\Phi^\dagger = \frac{\mathcal{C}}{N}I_{N\times N}$, and $\|\Phi\|^2 = \frac{\mathcal{C}}{N}$. This property makes it possible to achieve resilience to Gaussian noise in both the data and measurement domains.

Lemma 11. *Let ς be a vector with $\mathcal{C}$ iid $\mathcal{N}(0,\sigma_d^2)$ entries and e be a vector with N iid $\mathcal{N}(0,\sigma_m^2)$ entries. Let $\hbar = \Phi\varsigma$ and $u = \hbar + e$. Then u contains N entries, sampled iid from a Gaussian distribution $\mathcal{N}\left(0,\sigma^2\right)$, with $\sigma^2 = \frac{\mathcal{C}}{N}\sigma_d^2 + \sigma_m^2$.*

Proof: The tight frame property implies

$$\mathbb{E}\left[\hbar\hbar^\dagger\right] = E[\Phi\varsigma\varsigma^\dagger\Phi^\dagger] = \sigma_d^2\Phi\Phi^\dagger = \frac{\mathcal{C}}{N}\sigma_d^2 I.$$

[2] This assumption is only used in analyzing the performance of the algorithm in the stochastic noise model.

Therefore, $u = h + e$ can be considered to be white noise with variance σ^2. ■

2) Deterministic Noise Model: In the deterministic noise model, we shall assume that the magnitudes of the values of α are fixed, but their positions are distributed according to the model specified in Remark 1. Given a $DG(m,r)$ sieve, we now show that with constant probability $\|\Phi\left(\alpha - \alpha_{1\to k}\right)\|_2 \preceq \|\alpha - \alpha_{1\to k}\|_2$, and with probability $1 - \frac{k}{\mathcal{C}}$, $\|\Phi\left(\alpha - \alpha_{1\to k}\right)\|_2 \preceq \frac{\|\alpha - \alpha_{1\to k}\|_1}{\sqrt{k}}$.

Lemma 12. *Let Φ be a $DG(m,r)$ sieve, and let α be a signal in $\mathbb{C}^{\mathcal{C}}$ whose entries are distributed according to the model specified in Remark 1. If $0 \le \delta' \le 1$ then with probability $1 - \delta'$*

$$\|\Phi(\alpha - \alpha_{1\to k})\|_2 \le \frac{1}{\sqrt{\delta'}} \left(\|\alpha - \alpha_{1\to k}\|_2 + \frac{\|\alpha - \alpha_{1\to k}\|_1}{\sqrt{\mathcal{C} - 1}} \right). \tag{5}$$

Proof: We have

$$\|\Phi(\alpha - \alpha_{1\to k})\|^2 = \sum_{i=k+1}^{\mathcal{C}} |\alpha_i|^2 + \sum_{\substack{i,j \ge k+1 \\ i \ne j}} \alpha_i \overline{\alpha_j} \varphi_{\pi_i}^\dagger \varphi_{\pi_j}.$$

Given linearity of expectation we can rewrite this quantity in terms of average coherence and then apply Lemma 5 to obtain

$$\begin{aligned} &\mathbb{E}_\pi \left[\|\Phi(\alpha - \alpha_{1\to k})\|^2 \right] \\ &= \|\alpha - \alpha_{1\to k}\|^2 + \sum_{\substack{i,j \ge k+1 \\ i \ne j}} \alpha_i \overline{\alpha_j} \mathbb{E}_\pi \left[\varphi_{\pi_i}^\dagger \varphi_{\pi_j} \right] \\ &\le \left(\|\alpha - \alpha_{1\to k}\|_2 + \frac{\|\alpha - \alpha_{1\to k}\|_1}{\sqrt{\mathcal{C} - 1}} \right)^2. \end{aligned}$$

It follows from the Markov inequality that if $0 \le \delta' \le 1$ then with probability at least $1 - \delta'$

$$\|\Phi(\alpha - \alpha_{1\to k})\|^2 \le \frac{1}{\delta'} \left(\|\alpha - \alpha_{1\to k}\|_2 + \frac{\|\alpha - \alpha_{1\to k}\|_1}{\sqrt{\mathcal{C} - 1}} \right)^2. \tag{6}$$

■

IV. StRIP Functions

The Statistical Restricted isometry Property (StRIP) is defined by Calderbank et. al [13]. In this section we tighten the results in [13] for the Reed-Muller sieve. In addition, we generalize the StRIP notion to arbitrary functions $h : [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$, and arbitrary families of functions $h : [\mathcal{C}]^t \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$. We then use the StRIP property as a main tool for analyzing the sparse reconstruction algorithm in Section §VI:

Definition 13 (**(k,ϵ,δ)-StRIP**). *Let $\pi \doteq \{\pi_1, \cdots, \pi_{\mathcal{C}}\}$ be a random permutation of $\{1, \cdots, \mathcal{C}\}$. Let α be a k-sparse vector with support $\{\pi_1, \cdots, \pi_k\}$ and with fixed values $\alpha_1, \cdots, \alpha_k$. A function $h : [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ is (k,ϵ,δ)-StRIP if with probability $1 - \delta$ over the choice of π, the following two conditions are satisfied:*

1) *For every index $1 \le i \le k$,*

$$\left| \sum_{j : j \ne i} \alpha_j h(\pi_i, \pi_j) \right| \le \epsilon \|\alpha\|_2. \tag{7}$$

2) *For every index $w \in [\mathcal{C}] - \pi_{1\to k}$,*

$$\left| \sum_{j=1}^{k} \alpha_j h(w, \pi_j) \right| \le \epsilon \|\alpha\|_2. \tag{8}$$

A family of functions $\{h_t\}_{t=1}^k$, where $h_t : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$, is (k,ϵ,δ)-StRIP if with probability $1 - \delta$ over the choice of π, the following condition is satisfied:

1) *For every index $1 \le t \le k$*

$$\left| \sum_{j > t} \alpha_j h_t(\pi_{1\to t-1}, \pi_t, \pi_j) \right| \le \epsilon \|\alpha\|_2 \tag{9}$$

Remark 14. An $N \times \mathcal{C}$ measurement matrix satisfies the (k,ϵ,δ)-StRIP, if the function $h(i,j) = \varphi_i^\dagger \varphi_j$ is (k,ϵ,δ) StRIP. In other words, Definition 13 generalizes the definition of a StRIP for sensing matrices provided in [13], [25].

Similarly, we generalize the notion of coherence to cover arbitrary functions $h : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$, where $1 \le t \le k$.

Definition 15 (**(η,γ,N)-StRIP-able**). *A function $h : [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ is (η,γ,N)-StRIP-able if the following three conditions hold:*

- *(St1).*
$$\mu \doteq \max_{i \ne j} |h(\pi_i, \pi_j)| \le N^{-\eta}.$$
- *(St2).*
$$\nu \doteq \max_{i} |\mathbb{E}_{j \ne i} [h(\pi_i, \pi_j)]| \le N^{-\gamma}.$$
- *(St3). h is skew-symmetric. That is,*
$$h(x,y) = \overline{h(y,x)}$$
for all $x, y \in [\mathcal{C}]$.

For $1 \le t \le k$, a function $h : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ is (η,γ,N)-StRIP-able if for any fixed $\pi_{1\to t-1}$ the following three conditions hold:

- *(St1). For $1 \le j \le k$*
$$\mu \doteq \max_{\substack{w > t-1 \\ j \ne w}} |h(\pi_{1\to t-1}, \pi_w, \pi_j)| \le N^{-\eta}.$$
- *(St2). For $1 \le j \le k$*
$$\nu \doteq \max_{w > t-1} \left| \mathbb{E}_{\substack{j > t-1 \\ j \ne w}} [h(\pi_{1\to t-1}, \pi_w, \pi_j)] \right| \le N^{-\gamma}.$$
- *(St3). h is skew-symmetric in the last two variables. That is,*
$$h(\pi_{1\to t-1}, x, y) = \overline{h(\pi_{1\to t-1}, y, x)}$$

for all $x, y \in [\mathcal{C}]$.

In the following two theorems we show that StRIP-ability is a sufficient condition for satisfying the StRIP property.

Theorem 16. *Let $h : [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ be an (η, γ, N)-StRIP-able function. Then for all positive ϵ and for all $k \leq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, \epsilon^2 N^{2\gamma}\right\}$, h is $(k, 2\epsilon, \delta)$ StRIP with $\delta \leq 4\mathcal{C} \exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\}$.*

Theorem 17. *Let $H := \{h_t\}_{t=1}^{k}$ be a family of functions with $h_t : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ an (η, ∞, N)-StRIP-able function for each t. Then for all positive ϵ and for all $k \leq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, \epsilon^2 N^{2\gamma}\right\}$, H is (k, ϵ, δ)-StRIP with $\delta \leq 4\,k \exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\}$.*

The proofs of Theorem 16 and Theorem 17 are similar to the proof of Lemma 2 in [25], and are provided in Appendix §B. [3]

Corollary 18. *Let h be a (η, γ, N)-StRIP-able function. Given $\epsilon > 0$, let $k \leq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, \epsilon^2 N^{2\gamma}\right\}$. Suppose α is a k-sparse vector with uniformly random support and with fixed values $\alpha_1, \cdots, \alpha_k$. Then with probability at least $1 - 4\mathcal{C} \exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\}$,*

$$\left| \sum_{\substack{i,j \\ i \neq j}} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j) \right| \leq 2\sqrt{k}\epsilon \|\alpha\|^2.$$

Proof: The argument follows from Theorem 17, by applying the Cauchy-Schwarz inequality. ■

Next we show that if the sparsity level is sufficiently small, then with overwhelming probability

$$\left| \sum_i \alpha_i \sum_{j : j \neq i} \overline{\alpha_j} h(\pi_i, \pi_j) \right| \preceq \frac{\log \mathcal{C}}{N^\eta} \|\alpha\|^2 :$$

Theorem 19. *Let h be an (η, γ, N)-StRIP-able function from $[\mathcal{C}] \times [\mathcal{C}]$ to $\mathbb{C}$. Let π be a random permutation of $[\mathcal{C}]$, and let α be a k-sparse vectors with fixed values $\alpha_1, \cdots, \alpha_k$, such that $\operatorname{Supp}(\alpha) = \{\pi_1, \cdots, \pi_k\}$. Let $\xi > 0$ such that $k \leq \min\left\{\sqrt{C} - \frac{1}{2}, \xi N^\gamma\right\}$. Then for any $\epsilon > 0$*

$$\begin{aligned} &\Pr_{\pi}\left[\left| \sum_{i=1}^{k} \alpha_i \sum_{\substack{j=1 \\ j \neq i}}^{k} \overline{\alpha}_j h\left(\pi_i, \pi_j\right) \right| \geq 2\xi \|\alpha\|^2 \right] \\ &\leq 4 \exp\left\{ \frac{-\xi^2}{8192\epsilon^2} \right\} \end{aligned}$$

Proof: The proof of Theorem 19 is provided in Appendix §C. ■

Remark 20. *By setting $\varepsilon = O\left(\frac{\sqrt{\log \mathcal{C}}}{N^\eta}\right)$ and $\xi = O\left(\frac{\log \mathcal{C}}{N^\eta}\right)$ we can make sure that as long as $k \preceq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, N^{\gamma - \eta}\right\}$ the probability that $\left|\left(\sum_{i=1}^{k} \alpha_i \sum_{j=1}^{k} \overline{\alpha}_j h\left(\pi_i, \pi_j\right)\right) - \|\alpha\|^2\right| \succeq \frac{\log \mathcal{C}}{N^\eta} \|\alpha\|^2$ is $O\left(\frac{1}{Poly(\mathcal{C})}\right)$.*

V. The Witness-Averaging Algorithm

Algorithm 1 The Witness-Averaging Algorithm

```
for every witness $a \in \mathbb{F}_2^m$ do
    Pointwise multiply $f$ with a shifted version of itself (Auto-correlation).
    Compute the fast Hadamard transform: $\Gamma_a^{\ell}(f)$ (Power spectrum).
    Calculate $\Lambda_{\Delta,a} \doteq \imath^{-aP_\Delta a^\top} \Gamma_a^{a P_\Delta}(f)$ for every $\Delta \in [\mathcal{C}]$. (Demodulation).
end for
For each index $\Delta \in [\mathcal{C}]$, take the average of $\Lambda_\Delta \doteq \mathbb{E}_a[\Lambda_{\Delta,a}]$ over all $\Lambda_{\Delta,a}$ (Witness Averaging).
Let S be the position of the $k$ highest (in magnitude) average peaks (Thresholding).
Output $\hat{\alpha} = (\Phi_{\mathrm{S}}^\dagger \Phi_{\mathrm{S}})^{-1} \Phi_{\mathrm{S}}^\dagger f$ (Regression).
```

In this section we propose the Witness-Averaging Algorithm for sparse recovery from measurements obtained via the Reed-Muller sieve. The pseudocode for the Witness-Averaging Algorithm is shown in Algorithm 1. The algorithm identifies the signal model by analyzing the power spectrum of the pointwise product of the superposition f with a shifted version of itself. The Walsh-Hadamard transform of this pointwise product is the superposition of k Walsh tones and a background signal produced by cross-correlations between the k significant entries and cross-correlations between these k entries and the noise in the data domain and in the measurement domain. We shall prove that with high probability the energy in this background signal is uniformly distributed across the Walsh-Hadamard bins, and that this background bin energy is sufficiently small to enable threshold detection of the k tones. We show that sparse reconstruction is possible for $k = \tilde{O}(N)$ by averaging over all possible shifts.

The original quadratic reconstruction algorithm, proposed by Howard et. al [33], is a repurposing of the chirp detection algorithm commonly used in navigation radars which is known to work extremely well in the presence of noise. That algorithm has minimal complexity $k\,N\,\log^2 N$. When the algorithm is applied to $DG(m, r)$ frames [22], and two

[3] To prove the theorems, we start with a StRIP-able function of 2 arguments then we produce a StRIP-able function of 3 arguments by creating a "virtual" function of 3 arguments.

entries of the signal fall in the same orthonormal basis, the corresponding cross-term is a spurious Walsh-Hadamard tone which we refer to as an alias. Witnesses can be selected to distinguish these aliases but it is harder to analyze the algorithm, and reconstruction is only guaranteed for $k = \tilde{O}\left(\sqrt{N}\right)$ (See [13]). Here by using $DG(m,r)$ sieves, we guarantee that the *aliasing problem* never happens. This means that no two columns of Φ involve the same matrix P. This is crucial in the concentration analysis of the chirp-like terms.

The Witness Averaging Algorithm uses 2^m witnesses. Therefore, the starting loop of the algorithm takes $O\left(N(N + N\log N + \mathcal{C})\right) = O(N\mathcal{C})$ running time. The algorithm also needs $O(N\mathcal{C})$ time to calculate the averages (Step 6), $O(\mathcal{C}\log\mathcal{C})$ to find the k peaks via sorting, and $O(k^3)$ to calculate the pseudo-inverse. Therefore, the overall running-time of the algorithm is $O(N\mathcal{C})$. Performance bounds for recovery in the deterministic and stochastic noise models are given in Theorems 21 and 22.

Theorem 21. *Let α be a signal in $\mathbb{C}^{\mathcal{C}}$ taken from the average case signal analysis model (Remark 1). Then if*

1) $k \preceq \frac{\mathrm{MAR}}{m}\frac{N^{1-\frac{2r}{m}}}{\log\mathcal{C}}$, *and*
2) $\|u\|_2 \preceq \min\left\{\frac{1}{10}, \sqrt{\frac{N}{k\,\log\mathcal{C}}\,\mathrm{MAR}}\right\}\|\alpha\|_{\min}$,

then with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, Algorithm 1 successfully recovers the positions of the k largest entries of α. Moreover for every positive δ', with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)-\delta'$,

$$\|\hat{\alpha}-\alpha_{1\to k}\|_2 \leq 2\sqrt{2}\|e\|_2 + \sqrt{\frac{\mathcal{C}}{\mathcal{C}-1}\frac{4\sqrt{2}}{\sqrt{\delta'}}}\|\alpha-\alpha_{1\to k}\|_2.$$

Theorem 22. *Let ς be a vector with $\mathcal{C}$ iid $\mathcal{N}(0,\sigma_d^2)$ entries and e be a vector with N iid $\mathcal{N}(0,\sigma_m^2)$ entries. Define $\sigma \doteq \sqrt{\sigma_m^2 + \frac{\mathcal{C}}{N}\sigma_d^2}$. Let $\alpha_{1\to k}$ be a k-sparse signal with uniformly random support, and let $f = \Phi(\alpha_{1\to k}+\varsigma)+e$. Then if*

1) $k \preceq \frac{\mathrm{MAR}}{m}\frac{N^{1-\frac{2r}{m}}}{\log\mathcal{C}}$, *and*
2) $\sigma \preceq \min\left\{\frac{1}{10\log\mathcal{C}}, \sqrt{\frac{N^{1-\frac{r}{2m}}}{k\,\log^3\mathcal{C}}\,\mathrm{MAR}}\right\}\|\alpha\|_{\min}$,

then with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, Algorithm 1 successfully recovers the positions of the k largest entries of α. Moreover for all $0 \leq \varepsilon \leq \frac{1}{2}$

$$\Pr\left[\|\hat{\alpha}-\alpha_{1\to k}\|_2 \geq \sqrt{2(1+\varepsilon)k\left(\sigma_m^2+\frac{\mathcal{C}}{N}\sigma_d^2\right)}\right] \preceq O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)+\exp\left\{\frac{-3k\varepsilon^2}{16}\right\}.$$

In order to prove Theorems 21 and 22, we shall analyze the constituent steps of Algorithm 1 step by step. Let $f = y+u$, where $y = \Phi\alpha_{1\to k}$ and $u = e + \Phi\left(\alpha-\alpha_{1\to k}\right)$. By pointwise multiplication of f with a shifted version of itself, followed by the fast Hadamard transform, we form the power spectrum across all N Hadamard bins. Each bin ℓ then has the value

$$\Gamma_a^\ell(f) \doteq \frac{1}{\sqrt{N}}\sum_{x=1}^{N}(-1)^{\ell x^\top} f(x+a)\overline{f(x)} \tag{10}$$

Given the offset a, evidence for the presence or absence of a signal at position Δ in the data domain resides in the Hadamard bin $\ell = a\,P_\Delta$. Then at the Demodulation Step of Algorithm 1, for each index Δ in $[\mathcal{C}]$, we choose the Hadamard bin $\ell = a\,P_\Delta$ with a proper alignment shift in frequency. That is, we calculate the term $\Lambda_{\Delta,a}(f) = \imath^{-aP_\Delta a^\top}\Gamma_a^{aP_\Delta}(f)$. After aligning the phase, the final step is averaging over all offsets a. The notation $\mathbb{E}_a$ emphasizes that the average is taken over all offsets.

When an entry is present the evidence accumulated by the algorithm adds constructively, and when it is absent, the evidence adds destructively. We identify the positions of the largest entries in the signal by averaging over all possible witnesses. Having recovered the support, Theorem 4 guarantees that we can approximate the signal by solving a regression program. Figure 1 illustrates the role of witness averaging in the chirp reconstruction algorithm.

VI. Proof of Theorems 21 and 22

Let $\Lambda_{\Delta,a}(f) \doteq \imath^{-aP_\Delta a^\top}\Gamma_a^{a\,P_\Delta}(f)$. We start by using the linearity of expectation to decompose $\mathbb{E}_a\left[\Lambda_{\Delta,a}(f)\right]$ into its four constituents as follows:

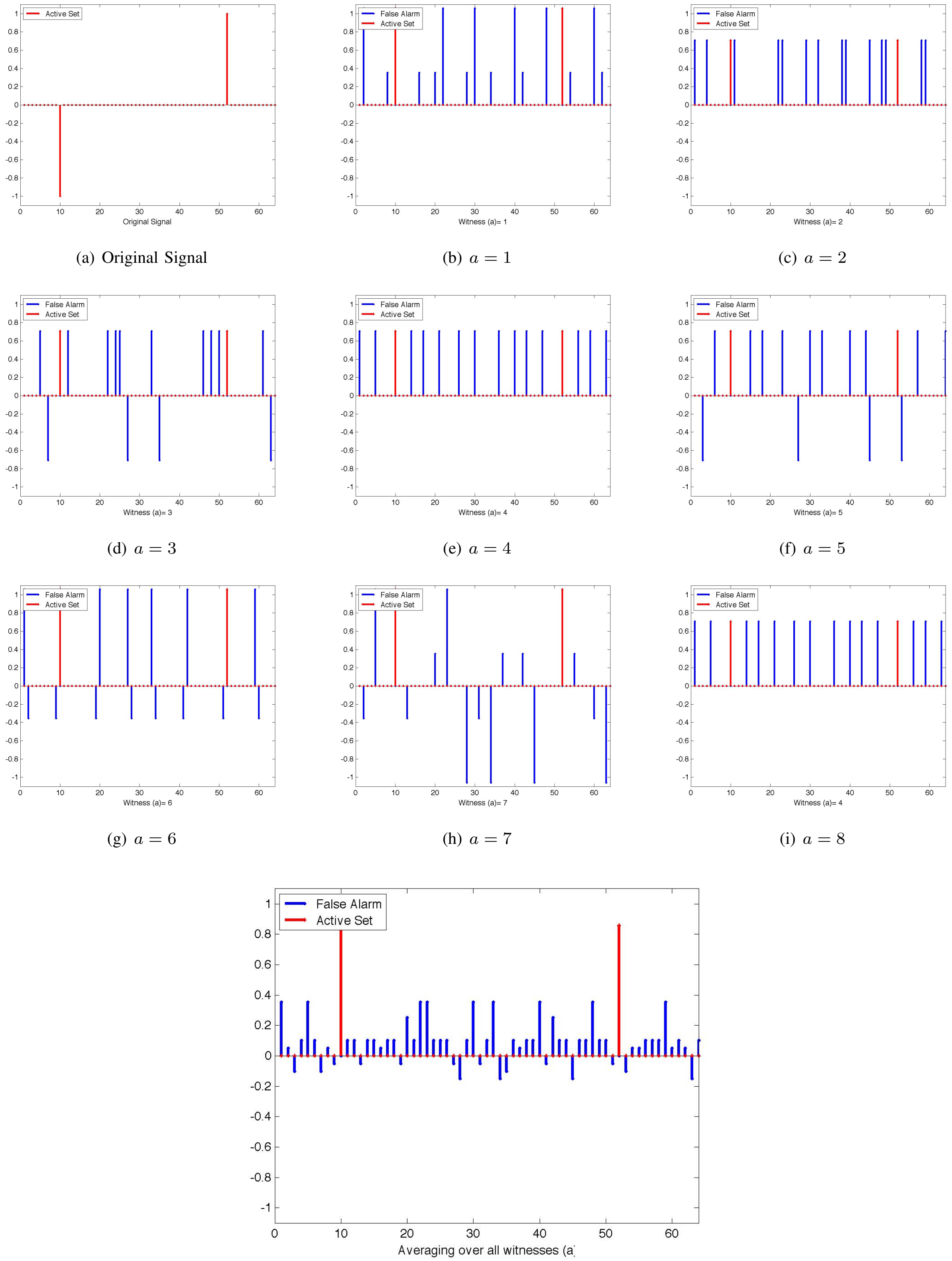
$$\begin{aligned}\mathbb{E}_a\left[\Lambda_{\Delta,a}(f)\right] &= \mathbb{E}_a\left[\Lambda_{\Delta,a}(y)\right] + \mathbb{E}_a\left[\Lambda_{\Delta,a}(u)\right] \\ &+ \mathbb{E}_a\left[\frac{\imath^{-aP_\Delta a^\top}}{\sqrt{N}}\left(\sum_x y(x+a)\overline{u(x)}(-1)^{aP_\Delta x^\top}\right)\right] \\ &+ \mathbb{E}_a\left[\frac{\imath^{-aP_\Delta a^\top}}{\sqrt{N}}\left(\sum_x \overline{y(x)}u(x+a)(-1)^{aP_\Delta x^\top}\right)\right],\end{aligned} \tag{11}$$

where $y = \Phi\alpha_{1\to k}$ and $u = f-y$. The following lemma shows that $\mathbb{E}_a\left[\Lambda_{\Delta,a}(y)\right]$ consists of k distinct Walsh tones staying on top of a uniform chirp-like residual term:

Lemma 23. *Let Φ be a $DG(m,r)$ sieve, and define $\delta_{a,i}^\ell \doteq \begin{cases} 1 & \text{if } a\,P_i+\ell = 0 \\ 0 & \text{otherwise}\end{cases}$. Then for all indices Δ in $[\mathcal{C}]$*

$$\mathbb{E}_a\left[\Lambda_{\Delta,a}(y)\right] = \sum_{i=1}^{k}\frac{|\alpha_i|^2}{\sqrt{N}}\delta_{\Delta,\pi_i} + \mathcal{R}_\Delta(y),$$

where $\mathcal{R}_\Delta(y)$ contains the demodulated chirp-like cross-



(a) Original Signal

(b) $a = 1$

(c) $a = 2$

(d) $a = 3$

(e) $a = 4$

(f) $a = 5$

(g) $a = 6$

(h) $a = 7$

(i) $a = 8$

(j) Averaging over all witnesses

Fig. 1. The role of witness averaging in the chirp reconstruction algorithm. Here α is a 2-sparse signal measured by a $DG(3, 1)$ sieve. For each witness a, the demodulated power spectrum $\Lambda_{\Delta,a}$ is plotted. Figure 1(j) shows the result of averaging over all witnesses. When a signal is present, the evidence adds constructively, and when it is not present the evidence adds destructively.

terms:

$$\sum_{\substack{i\in\{1,\cdots,k\}\\ \pi_i\neq\Delta}} \frac{|\alpha_i|^2}{\sqrt{N}}\mathbb{E}_a\left[i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}\delta_{a,\pi_i}^{aP_\Delta}\right] \tag{12}$$

$$+\frac{1}{N^{\frac{3}{2}}}\sum_{i=1}^{k}\sum_{j\neq i}\alpha_i\overline{\alpha_j}\sum_x i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}\mathcal{E}_x(\pi_i,\Delta),$$

with

$$\mathcal{E}_x(\pi_i,\Delta)\doteq\mathbb{E}_a\left[i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}(-1)^{\left(aP_{\pi_i}-aP_\Delta\right)x^\top}\right].$$

Proof: $\Gamma_a^\ell(y)$ is obtained by first multiplying y with a shifted version of itself, and then calculating the ℓ^{th} Hadamard transform coefficient. We have

$$y(x+a)\overline{y(x)}=\sum_{i=1}^{k}\alpha_i\varphi_{\pi_i}(x+a)\overline{\left(\sum_{j=1}^{k}\alpha_j\varphi_{\pi_j}(x)\right)}$$

$$=\sum_{i,j=1}^{k}\alpha_i\overline{\alpha_j}\varphi_{\pi_i}(x+a)\overline{\varphi_{\pi_j}(x)}.$$

It follows from the construction of the DG Sieve (Equation (1)) that

$$\varphi_{\pi_i}(x+a)\overline{\varphi_{\pi_j}(x)}=\frac{1}{N}i^{aP_{\pi_i}a^\top}(-1)^{aP_{\pi_i}x^\top}i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}.$$

In other words, auto-correlating a pure tone φ_{π_i} with a shifted version of itself generates a Walsh tone. Now using the fast Hadamard transform

$$\Gamma_a^\ell(y)=\sum_{i=1}^{k}\frac{i^{aP_{\pi_i}a^\top}|\alpha_i|^2}{\sqrt{N}}\delta_{a,\pi_i}^{\ell} \tag{13}$$

$$+\frac{1}{N^{\frac{3}{2}}}\sum_{i=1}^{k}\sum_{j\neq i}\alpha_i\overline{\alpha_j}i^{aP_{\pi_i}a^\top}\sum_x(-1)^{\left(aP_{\pi_i}+\ell\right)x^\top}i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}.$$

By demodulating Equation (13) (forming $\Lambda_{\Delta,a}$), and then averaging over all choices of a we get

$$\mathbb{E}_a\left[\Lambda_{\Delta,a}(y)\right]=\sum_{i=1}^{k}\mathbb{E}_a\left[\frac{i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}|\alpha_i|^2}{\sqrt{N}}\delta_{a,\pi_i}^{aP_\Delta}\right]$$

$$+\frac{1}{N^{\frac{3}{2}}}\sum_{i=1}^{k}\sum_{j\neq i}\alpha_i\overline{\alpha_j}\sum_x i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}\mathcal{E}_x(\pi_i,\Delta).$$

The first term can further be expanded as

$$\sum_{i=1}^{k}\mathbb{E}_a\left[\frac{i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}|\alpha_i|^2}{\sqrt{N}}\delta_{a,\pi_i}^{aP_\Delta}\right]$$

$$=\sum_{i=1}^{k}\mathbb{E}_a\left[\frac{i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}|\alpha_i|^2}{\sqrt{N}}\delta_{\pi_i,\Delta}\right]$$

$$+\sum_{i=1}^{k}\mathbb{E}_a\left[\frac{i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}|\alpha_i|^2}{\sqrt{N}}\delta_{a,\pi_i}^{aP_\Delta}(1-\delta_{\pi_i,\Delta})\right]=$$

$$\sum_{i=1}^{k}\left(\frac{|\alpha_i|^2}{\sqrt{N}}\delta_{\pi_i,\Delta}+\frac{|\alpha_i|^2}{\sqrt{N}}(1-\delta_{\pi_i,\Delta})\mathbb{E}_a\left[i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}\delta_{a,\pi_i}^{aP_\Delta}\right]\right)$$

■

The Walsh-Hadamard tones appear as exactly k spikes $\frac{1}{\sqrt{N}}\sum_{i=1}^{k}|\alpha_i|^2$ above a constant background signal $\mathcal{R}_\Delta(y)$. In the rest of this section, we use MAR and $\|\alpha\|_{\min}$ to abbreviate $\mathrm{MAR}(\alpha_{1\rightarrow k})$ and $\|\alpha_{1\rightarrow k}\|_{\min}$. In the following lemma we show that with overwhelming probability, for every index $\Delta\in[\mathcal{C}]$, the background chirp-like terms have magnitude at most $\frac{m\ \log\mathcal{C}}{N^{\frac{3}{2}-\frac{2r}{m}}}\|\alpha_{1\rightarrow k}\|^2$.

Lemma 24. *Let $k\preceq\mathrm{MAR}\frac{N^{1-\frac{2r}{m}}}{\log\mathcal{C}}$, then with probability $1-O\left(\frac{1}{Poly(\mathcal{C})}\right)$ for every index Δ in $[\mathcal{C}]$: $\mathcal{R}_\Delta(y)\preceq\frac{m\log\mathcal{C}\,\|\alpha_{1\rightarrow k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}}+\sum_{i=1}^{k}\frac{|\alpha_i|^2}{10N^{\frac{1}{2}}}\delta_{\Delta,\pi_i}$.*

Proof: The proof of Lemma 24 is provided in Appendix §D. ■

Next we show that the cross-correlation between the signal and the noise also provides uniform background terms with sufficiently small magnitudes:

Lemma 25. *Let u denote the total noise vector. If $\|\alpha\|_{\min}\geq 10\|u\|_2$ then with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, for every index Δ in $[\mathcal{C}]$:*

$$\left|\mathbb{E}_a\left[\frac{i^{-aP_\Delta a^\top}}{\sqrt{N}}\left(\sum_x y(x+a)\overline{u(x)}(-1)^{aP_\Delta x^\top}\right)\right]\right|$$

$$\preceq\frac{\sqrt{\log\mathcal{C}}\|u\|_2\|\alpha\|_2}{N}+\sum_{i=1}^{k}\frac{|\alpha_i|^2}{10N^{\frac{1}{2}}}\delta_{\Delta,\pi_i}.$$

Moreover, if the elements of u have iid random signs, then the requirement on the noise magnitude can be relaxed to $\|\alpha\|_{\min}\geq\frac{10\sqrt{\log\mathcal{C}}\|u\|_2}{\sqrt{N}}$, and with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$ for every Δ

$$\left|\mathbb{E}_a\left[\frac{i^{-aP_\Delta a^\top}}{\sqrt{N}}\left(\sum_x y(x+a)\overline{u(x)}(-1)^{aP_\Delta x^\top}\right)\right]\right|$$

$$\preceq\frac{\log\mathcal{C}\|\alpha\|_2\|u\|_2}{N^{\frac{3}{2}-\frac{r}{m}}}+\sum_{i=1}^{k}\frac{|\alpha_i|^2}{10N^{\frac{1}{2}}}\delta_{\Delta,\pi_i}.$$

The proof of Lemma 25 is provided in Appendix §E.

The term $\mathbb{E}_a\left[\frac{i^{-aP_\Delta a^\top}}{\sqrt{N}}\left(\sum_x\overline{y(x)}u(x+a)(-1)^{aP_\Delta x^\top}\right)\right]$ can be bounded similarly. The Cauchy-Schwarz inequality can be used to bound $\mathbb{E}_a\left[\Lambda_{\Delta,a}(u)\right]$ by $\frac{1}{\sqrt{N}}\|u\|^2$ which is negligible comparing to the other three terms. Here we have shown that the chirp-like cross-terms, and the cross correlation of signal with noise are distributed uniformly across all indices Δ. Hence, by thresholding $\mathbb{E}_a\left[\Lambda_{\Delta,a}(f)\right]$ we can recover the support of $\alpha_{1\rightarrow k}$:

Lemma 26. *If*

1) $k\preceq\frac{\mathrm{MAR}}{m}\frac{N^{1-\frac{2r}{m}}}{\log\mathcal{C}}$, *and*
2) $\|u\|_2\preceq\min\left\{\frac{1}{10},\sqrt{\frac{N}{k\ \log\mathcal{C}}\mathrm{MAR}}\right\}\|\alpha\|_{\min}$,

then with probability $1 - O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$*, chirp reconstruction successfully recovers the positions of the* k *significant entries of* $\alpha_{1\to k}$*.*

Proof: Chirp detection generates k Walsh tones with magnitudes at least $\frac{\|\alpha\|_{\min}^2}{\sqrt{N}}$ above a uniform background signal. Furthermore, there exist constants $\mathrm{c}_1, \mathrm{c}_2$ such that with probability at least $1 - O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$ every background signal at every index Δ is bounded by

$$\frac{3}{10}\sum_{i=1}^{k}\frac{|\alpha_i|^2}{N^{\frac{1}{2}}}\delta_{\Delta,\pi_i} + \frac{\mathrm{c}_1 m\log\mathcal{C}\,\|\alpha_{1\to k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}} + \frac{\mathrm{c}_2\sqrt{\log\mathcal{C}}\|u\|_2\|\alpha_{1\to k}\|_2}{N}.$$

Hence, if $\frac{\mathrm{c}_1 m\log\mathcal{C}\,\|\alpha_{1\to k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}} + \frac{\mathrm{c}_2\sqrt{\log\mathcal{C}}\|u\|_2\|\alpha\|_2}{N}$ is smaller than $\frac{7\|\alpha\|_{\min}^2}{20\sqrt{N}}$ then the k tones pop up and we can detect them by thresholding. Therefore, it is sufficient to ensure that $\frac{m\log\mathcal{C}\,\|\alpha_{1\to k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}} \preceq \frac{\|\alpha\|_{\min}^2}{N^{\frac{1}{2}}}$, and $\frac{\sqrt{\log\mathcal{C}}\|u\|_2\|\alpha\|_2}{N} \preceq \frac{\|\alpha\|_{\min}^2}{N^{\frac{1}{2}}}$. ■

The following lemma indicates that we can tolerate larger noise magnitude in the stochastic noise regime:

Lemma 27. *Suppose the elements of* u *have independent random signs. Then if*

1) $k \preceq \frac{\mathrm{MAR}}{m}\frac{N^{1-\frac{2r}{m}}}{\log\mathcal{C}}$*, and*
2) $\|u\|_2 \preceq \min\left\{\frac{\sqrt{N}}{10\sqrt{\log\mathcal{C}}}, \frac{N^{1-\frac{r}{m}}\sqrt{\mathrm{MAR}}}{\sqrt{k}\,\log\mathcal{C}}\right\}\|\alpha\|_{\min}$,

then with probability $1 - O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$*, chirp reconstruction successfully recovers the positions of the* k *significant entries of* $\alpha_{1\to k}$*.*

Proof: The proof is similar to the proof of Lemma 26. The only difference is that now we need to ensure that $\frac{m\log\mathcal{C}\,\|\alpha_{1\to k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}} \preceq \frac{\|\alpha\|_{\min}^2}{N^{\frac{1}{2}}}$, and $\frac{\log\mathcal{C}\|u\|_2\|\alpha\|_2}{N^{\frac{3}{2}-\frac{r}{m}}} \preceq \frac{\|\alpha\|_{\min}^2}{N^{\frac{1}{2}}}$. ■

Remark 28. *If the conditions of Lemma 26 or Lemma 27 holds, then we can perform the Thresholding Step of Algorithm 1 even without knowing the true model order* k*. The Thresholding Step can be performed by forming* $\mathbb{E}_a\left[\Lambda_{\Delta,a}(f)\right]$*, and collecting the indices* Δ *that have magnitudes* $|\mathbb{E}_a\left[\Gamma_\Delta(f)\right]|$ *larger than* $\frac{7\|\alpha\|_{\min}^2}{20\sqrt{N}}$*.*

Remark 29. *Chirp reconstruction is able to detect the presence or absence of a signal at any given index* Δ *in the data domain without needing to first reconstruct the entire signal. The complexity of detection is* $O\left(N^2\log N\right)$ *(we have* 2^m *witnesses, and for each witness the bottleneck is the* $O(N\log N)$ *time of calculating the fast Hadamard transform). If the signal* α *were the wavelet decomposition of an image, then chirp reconstruction can be applied to the measured signal to recover thumbnails and to zoom in on areas of interest.*

Having identified the support, we now analyze the sparse reconstruction guarantees of Algorithm 1.

Lemma 30. *Let* $S = \{\pi_1, \cdots, \pi_k\}$ *and let* $\hat{\alpha} \doteq \arg\min_{\alpha^+}\|f - \Phi_S\alpha^+\|^2$*. Then for every positive* δ'*, with probability* $1-\delta'$

$$\|\Phi\left(\hat{\alpha} - \alpha_{1\to k}\right)\|_2 \leq 2\|e\|_2 + \sqrt{\frac{\mathcal{C}}{\mathcal{C}-1}}\frac{4}{\sqrt{\delta'}}\|\alpha - \alpha_{1\to k}\|_2.$$

Moreover, if the data domain noise consists of $\mathcal{C}$ *iid* $\mathcal{N}\left(0, \sigma_d^2\right)$ *random variables, and the measurement noise contains* N *iid* $\mathcal{N}\left(0, \sigma_m^2\right)$ *random variables, then for all* $0 \leq \varepsilon \leq \frac{1}{2}$

$$\Pr\left[\|\Phi\left(\hat{\alpha} - \alpha_{1\to k}\right)\|_2 \geq \sqrt{(1+\varepsilon)k\left(\sigma_m^2 + \frac{\mathcal{C}}{N}\sigma_d^2\right)}\right] \leq \exp\left\{\frac{-3k\varepsilon^2}{16}\right\}.$$

Proof: It follows from Lemma 12 that for every positive δ', with probability at least $1-\delta'$,

$$\begin{aligned}&\|\Phi\left(\hat{\alpha} - \alpha_{1\to k}\right)\|_2\\ &\leq \|f - \Phi\alpha_{1\to k}\|_2 + \|f - \Phi\hat{\alpha}\|_2 \leq 2\,\|f - \Phi\alpha_{1\to k}\|_2\\ &\leq 2\|e\|_2 + \frac{2}{\sqrt{\delta'}}\left(\|\alpha - \alpha_{1\to k}\|_2 + \frac{\|\alpha - \alpha_{1\to k}\|_1}{\sqrt{\mathcal{C}-1}}\right)\\ &< 2\|e\|_2 + \sqrt{\frac{\mathcal{C}}{\mathcal{C}-1}}\frac{4}{\sqrt{\delta'}}\|\alpha - \alpha_{1\to k}\|_2.\end{aligned}$$

In the stochastic noise regime, Lemma 11 states that u has N iid $\mathcal{N}\left(0, \frac{\mathcal{C}}{N}\sigma_d^2 + \sigma_m^2\right)$ elements. Then $\|\Phi\left(\hat{\alpha} - \alpha_{1\to k}\right)\|^2 = \|\mathcal{P}_S[u]\|^2$, where $\mathcal{P}_S[u]$ is the projection of u onto the space spanned by S. The result then follows from the concentration of the χ^2 distribution (Proposition 33). ■

Remark 31. *As long as* $k \leq \frac{c\,\mathcal{C}}{\|\Phi\|^2\log\mathcal{C}}$*, we can use Theorems 3 and 4 to translate the approximation error in the measurement domain to approximation error in the data domain. In particular with probability at least* $1 - \frac{2}{\mathcal{C}} - \delta'$*:*

$$\|\hat{\alpha} - \alpha_{1\to k}\|_2 \leq 2\sqrt{2}\|e\|_2 + \sqrt{\frac{\mathcal{C}}{\mathcal{C}-1}}\frac{4\sqrt{2}}{\sqrt{\delta'}}\|\alpha - \alpha_{1\to k}\|_2.$$

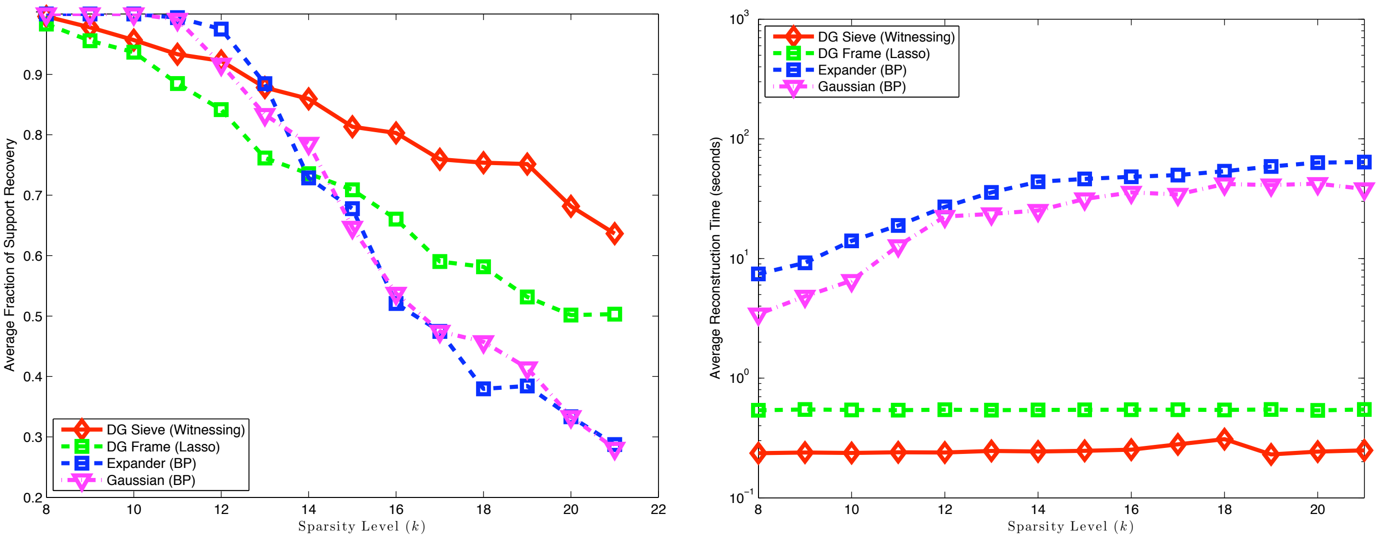
Moreover, if the noise is Gaussian, then

$$\Pr\left[\|\hat{\alpha} - \alpha_{1\to k}\|_2 \geq \sqrt{2(1+\varepsilon)k\left(\sigma_m^2 + \frac{\mathcal{C}}{N}\sigma_d^2\right)}\right] \leq \frac{2}{\mathcal{C}} + \exp\left\{\frac{-3k\varepsilon^2}{16}\right\}.$$

VII. Experimental Results

A. Recovering the Support of Random Sparse Signals

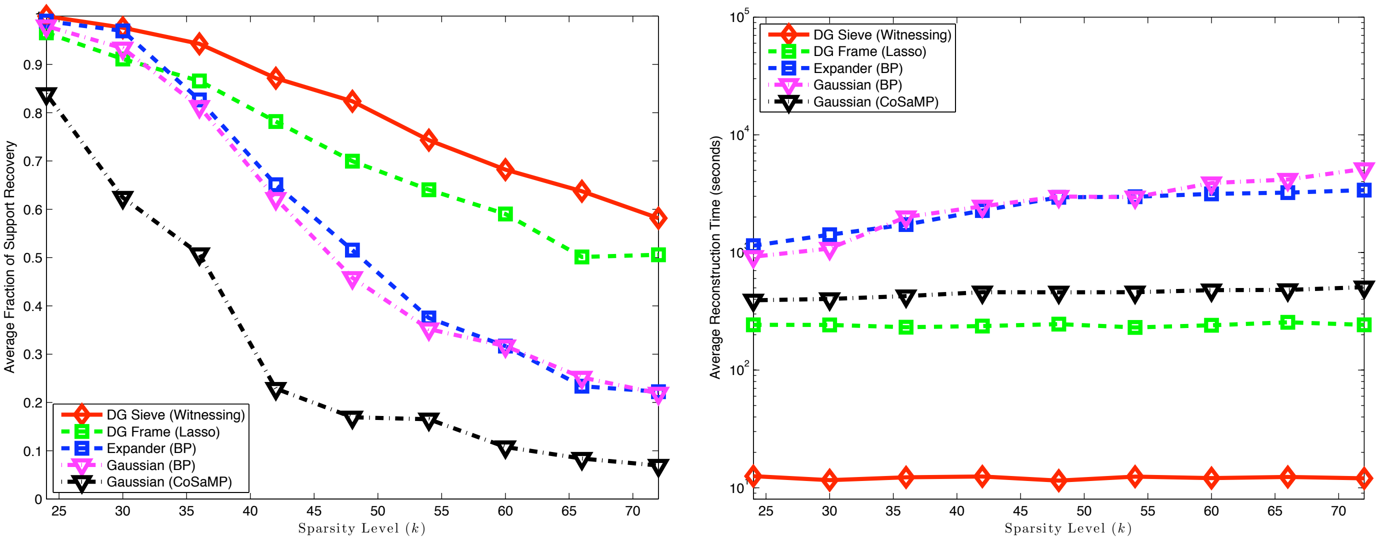
In this Section we present the results of numerical experiments that make it possible to compare different measurement matrices and different algorithms for sparse recovery and model selection. We examined DG sieves, DG frames,



(a) Average fraction of the support that is reconstructed successfully as a function of the sparsity level k.

(b) Average reconstruction time (logarithmic scale) in the noiseless regime for different sensing matrices.

Fig. 2. Comparison between the chirp reconstruction algorithm for Delsarte-Goethals sieve $DG(7, 1)$, with Basis Pursuit algorithm for Gaussian and Expander matrices of the same size, and with LASSO algorithm for Delsarte-Goethals frame $DG(7, 0)$.



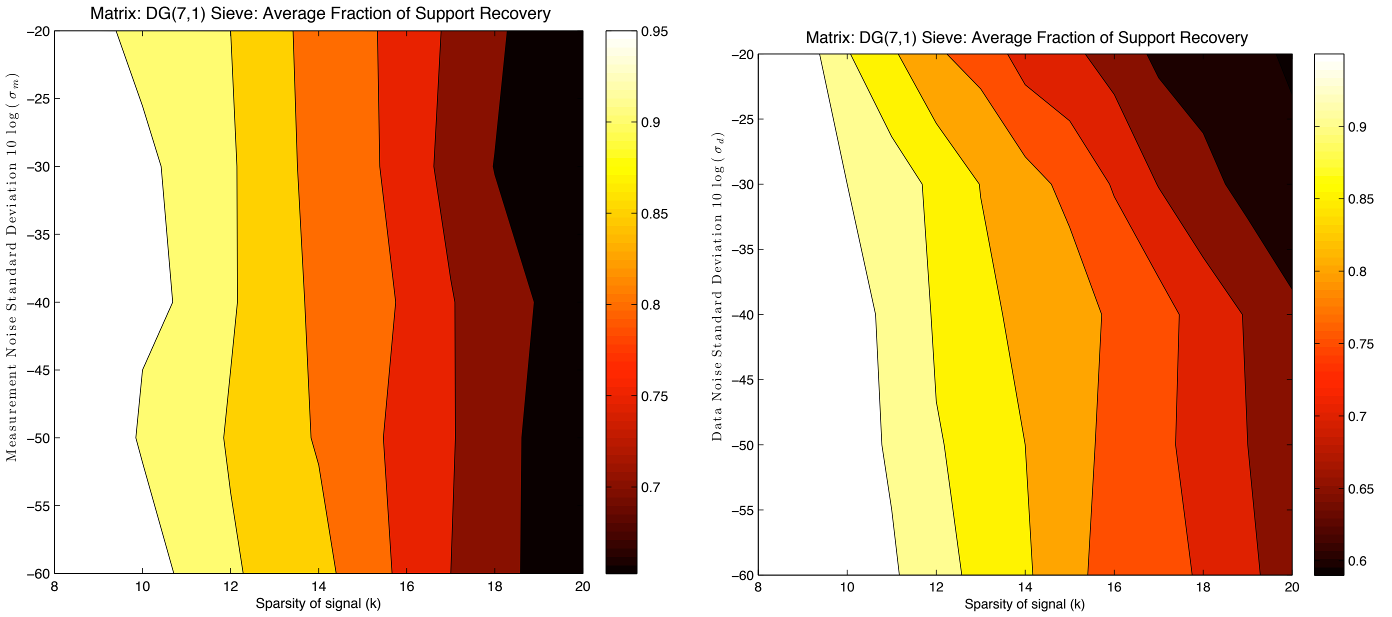
(a) Average fraction of the support that is reconstructed successfully as a function of the sparsity level k.

(b) Average reconstruction time (logarithmic scale) in the noiseless regime for different sensing matrices.

Fig. 3. Comparison between the chirp reconstruction algorithm for Delsarte-Goethals sieve $DG(9, 1)$, with Basis Pursuit algorithm for Gaussian and Expander matrices, CoSaMP algorithm, and with LASSO algorithm for Delsarte-Goethals frame $DG(9, 0)$.

Gaussian matrices, and Expander matrices. The DG frames are equiangular tight frames obtained by exponentiating all possible Delsarte Goethals codewords. If the sparse signal has random support and random sign, then the LASSO program will successfully recover the support [14], [22]. Expander graphs are bipartite graphs where the adjacency matrix satisfies a Restricted Isometry Property with respect to the ℓ_1 norm RIP-1); this condition guarantees that Basis Pursuit [1] will successfully recover signals that are sufficiently sparse [34]. Here we used the $\ell_1 - magic$ algorithm [35] to solve the Basis Pursuit program, and used the SpaRSA algorithm [36] to solve the LASSO program. The SpaRSA algorithm with ℓ_1 regularization parameter $\lambda = 10^{-9}$ was used for signal reconstruction in the noiseless case, and the parameter was adjusted to $2\sqrt{2 \log \mathcal{C}}\sigma$ in the noisy case [14]. The reason for using SpaRSA is that is designed to solve complex valued LASSO programs. We also present the results of numerical experiments with CoSaMP [7].

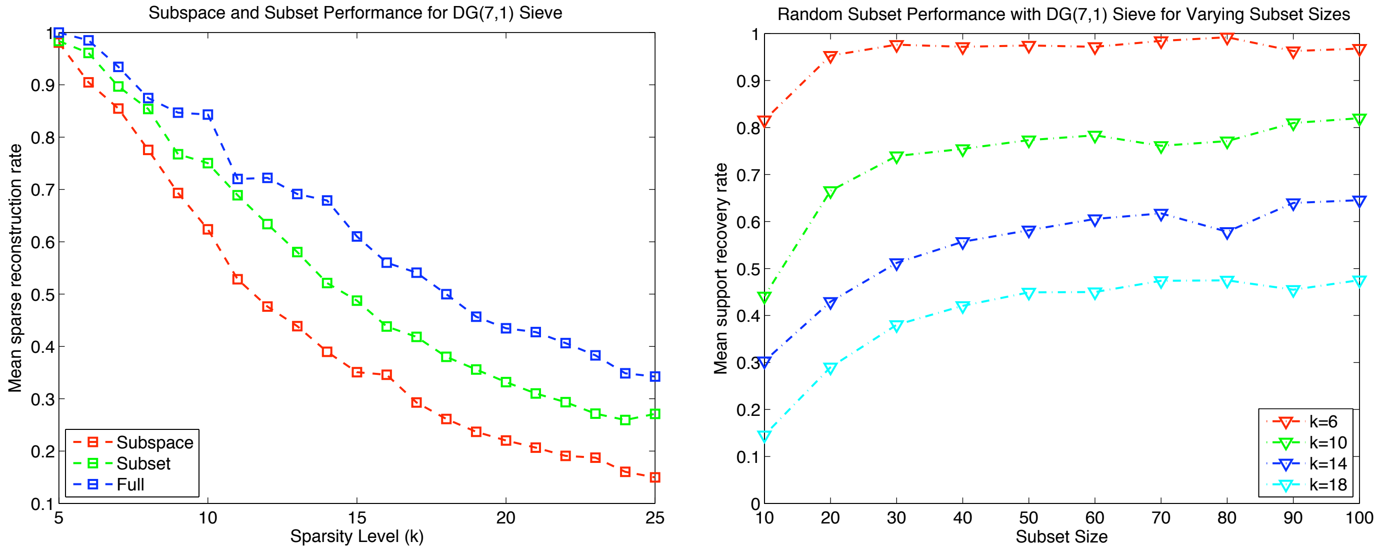
For Gaussian matrices, we sampled 10 iid random matrices



(a) The impact of the noise in the measurement domain on the accuracy of the sparse approximation for different sensing matrices.

(b) The impact of the noise in the data domain on the accuracy of the sparse approximation for different sensing matrices.

Fig. 4. The effect of the noise in the measurement domain (left), and in the data domain (right), on the performance of the chirp reconstruction algorithm.



(a) Subspace and subset performance for $DG(7,1)$ sieve.

(b) Random subset performance with $DG(7,1)$ for varying subset size.

Fig. 5. The impact of subsampling a random subspace/subset on the performance of the witness averaging algorithm with $DG(7,1)$ sieve.

independently to eliminate the exponentially small chance of getting a sample Φ not satisfying the RIP property, and the median of the results among all 10 random matrices was provided.

The experiments relate accuracy of sparse recovery to the sparsity level and the Signal to Noise Ratio (SNR). We measured the accuracy in terms of the statistical $0-1$ loss metric, capturing the fraction of signal support that is successfully recovered. Without loss of generality, we let each reconstruction algorithm output a k-sparse vector $\hat{\alpha}$. The statistical $0-1$ loss is the fraction of the support of α that is not recovered in $\hat{\alpha}$. Each experiment was repeated 2000 times, and the average $0-1$ loss was reported.

Figure 2 plots statistical $0-1$ loss and complexity (average reconstruction time) as a function of the sparsity level k. We generated k-sparse signals with uniformly random support, with random signs, and with the amplitude of non-zero entries set equal to 1. Four different sensing matrices are compared; a Gaussian matrix, an Expander Graph with left-degree $d=16$, a $DG(7,0)$ frame and a $DG(7,1)$ sieve. Figure 3 shows the results of applying the same experiments to larger sensing matrices.

Figure 4(a) plots statistical $0-1$ loss as a function of noise in the measurement domain and Figure 4(b) does the same for noise in the data domain. In the measurement noise study, a $\mathcal{N}(0, \sigma^2)$ iid measurement noise vector is added to the sensed vector to obtain the N dimensional vector f. We use a similar method to study noise in the data domain.

The contour plots show the average fraction of support that is recovered successfully as a function of the sparsity level k (horizontal axis), and the noise standard deviation $10\log_{10}(\sigma)$ (vertical axis). The sparsity level ranges between 8 and 20, and the noise standard deviation ranges between 10^{-6} to 10^{-2}. Recall that the witness averaging algorithm does not require independence among the signs of the elements of α. In this experiment the support was chosen uniformly at random and every element of α was non-negative[4]. Each experiment was then repeated 200 times. The average fraction of the successfully recovered support is illustrated by the intensity of the corresponding pixel in the contour plot.

B. Throwing out Witnesses: Random Subsets and Random Subspaces

A natural question to ask about the chirp reconstruction algorithm is the following: what is the effect of pruning the number of witnesses on the fidelity of support recovery? There are two distinct approaches to subsampling the set of all witnesses $\mathbb{F}_2^m$:

- Select an additive subspace of $\mathbb{F}_2^m$.
- Select a random subset of witnesses.

In experiments on both approaches, we used randomness to generate the subsets. In the first case, we only considered additive subspaces which were spanned by standard basis vectors. To generate such a subspace for a given dimension $t < m$, we chose t standard basis vectors uniformly at random from the m possible vectors. To generate a random subset of witnesses for a specified size $t < 2^m$, we chose t vectors uniformly at random from the 2^m possible choices. Figure 5(a) compares subspace and subset performance, with the performance of the full set of witnesses as a bench mark. In this experiment, we used the $DG(7,1)$ sieve, and for each k ran 1000 signals against 15 subspaces of dimension 5 (and thus size $2^5 = 32$), and also ran 1000 signals against 100 subsets of size 32. The signal generation process was the same as in the previous section.

It follows from Figure 5(a), and other similar experiments, that the subset subsampling strategy outperforms the subspace strategy. Moreover, both track the accuracy of the full witness set fairly closely, despite the fact that only $\frac{1}{4}$ of the witnesses are being used. This has promising implications for reducing runtime while maintaining accuracy.

Figure 5(b) shows the impact of the subset size on the support recovery rate. Each data point represents 400 signals run against 50 different random subsets of a given size. The signal generation process was the same as before. Note that as the subset size decreases, the support recovery rate remains fairly constant until around 40. In other words, we can get almost the same performance by selecting only a $\frac{1}{3}$ fraction of the witnesses uniformly at random, and discarding the rest of them.

VIII. Conclusion

In compressed sensing the entries of the measurement vector constitute evidence for the presence or absence of a signal at any given location in the data domain. We have shown that the Witness-Averaging Algorithm is able to identify the support set of most sparse vectors from measurements obtained via the Reed-Muller sieve. This model selection goal can be achieved without requiring that the signal entries be independent. We have also demonstrated feasibility of local decoding where attributes of the signal are deduced from the measurements without explicitly reconstructing the full signal.

Our reconstruction algorithms are resilient to noise. The average-case ℓ_2/ℓ_2 error bounds of the Witness-Averaging Algorithm is a direct consequence of the structured construction of the Reed-Muller sieve. This type of bounds are tighter than the ℓ_2/ℓ_1 bounds arising from random ensembles, and are information-theoretically impossible in the worst-case compressed sensing framework. Experimental results were also provided to support the fidelity of the proposed algorithm. Future directions involve generalizing the algorithm to work with other other structured matrices (e.g. Gabor frames, BCH matrices, etc), and analyzing the information-theoretic limits of average-case compressed sensing in general, and the Witness Averaging Algorithm in particular.

Acknowledgment

The authors would like to thank Waheed Bajwa, Ingrid Daubechies, Marco Duarte, Robert Schapire, and Karthik Sridharan for sharing many valuable insights. The work of R. Calderbank and S. Jafarpour is supported in part by NSF under grant DMS 0701226, by ONR under grant N00173-06-1-G006, and by AFOSR under grant FA9550-05-1-0443.

[4] We observed similar results for other sign patterns.

References

[1] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics, Vol. 59 (8) , pp. 1207-1223.*, 2006.

[2] D. Donoho, "Compressed Sensing," *IEEE transactions on Information Theory, Vol. 52 (4), pp. 1289-1306*, April 2006.

[3] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies," *IEEE Transactions on Information Theory, Vol. 52 (12), pp. 5406-5425*, December 2006.

[4] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE transactions on Information Theory, Vol. 52 (2), pp. 489-509*, 2006.

[5] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Asymptotic Achievability of the CramŕRao Bound for Noisy Compressive Sampling," *IEEE Trans. Signal Processing, Vol. 57(3), pp. 1233-1236*, March 2009.

[6] J. Tropp, "Greed is Good: Algorithmic Results for Sparse Approximation," *IEEE Transactions on Information Theory, Vol. 50 (10), pp. 2231-2242*, October 2004.

[7] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples.," *Applied and Computational Harmonic Analysis, Vol. 26 (3), pp. 301-321*, May 2009.

[8] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing: Closing the gap between performance and complexity," *IEEE Transactions on Information Theory, Vol. 55 (5), pp. 2230 - 2249*, 2009.

[9] A. Gilbert, M. Strauss, J. Tropp, and R. Vershynin, "One sketch for all: fast algorithms for compressed sensing," *Proceedings of the thirty-ninth annual ACM Symposium on Theory of Computing (STOC), pp. 237-246*, 2007.

[10] M. Akakaya, J. Park, and V. Tarokh, "A coding theory approach to noisy compressive sensing using low density frames," *submitted*, 2010.

[11] D. Baron, S. Sarvotham, and R. Baraniuk, "Bayesian Compressed Sensing via Belief Propagation," *Rice ECE Department Technical Report TREE 0601*, 2006.

[12] D. Donoho, A. Maleki, and A. Montanari, "The Noise-Sensitivity Phase Transition in Compressed Sensing," *Preprint*, 2010.

[13] R. Calderbank, S. Howard, and S. Jafarpour, "Construction of a large class of Matrices satisfying a Statistical Isometry Propery," *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Compressive Sensing, Vol. 4(2), pp. 358-374*, 2010.

[14] E. Candès and J. Plan, "Near-ideal model selection by ℓ_1 minimization," *Annals of Statistics, Vol. 37, pp. 2145-2177*, 2009.

[15] P. Indyk and M. Ruzic, "Near-optimal sparse recovery in the $\ell 1$ norm," *49th Annual IEEE Symposium on Foundations of Computer Science, 2008 (FOCS '08), pp. 199-207*, 2008.

[16] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient compressed Sensing using Optimized Expander Graphs," *IEEE Transactions on Information Theory, Vol. 55 (9), pp. 4299-4308.*, 2009.

[17] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of American Mathematical Society Vol. 22, pp. 211-231*, 2009.

[18] G. Cormode and S. Muthukrishnan, "Combinatorial algorithms for Compressed Sensing," *In Proceedings of 40th Annual Conference on Information Sciences and Systems (CISS), Princeton*, 2006.

[19] A. Gilbert, S. Guha, P. Indyk, M. Muthukrishnan, and M. Strauss, "Near-optimal sparse fourier representations via sampling," *Proceedings of the 34th annual ACM Symposium on Theory of Computing (STOC)*, 2002.

[20] M. Akcakaya and V. Tarokh, "A frame construction and a universal distortion bound for sparse representations," *IEEE Transactions on Signal Processing, Vol. 56 (6) , pp. 2443-2450*, June 2008.

[21] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Annals of Statistics, Vol. 35 (6), pp. 2313-2351*, 2007.

[22] R. Calderbank and S. Jafarpour, "Reed Muller Sensing Matrices and the LASSO," *International Conference on Sequences and their Applications (SETA), pp. 442463*, 2010.

[23] J. Bourgain, S. Dilworth, K. Ford, S. Konyagin, and D. Kutzarova, "Breaking the k^2 barrier for explicit rip matrices," *preprint*, 2010.

[24] W. Bajwa, J. Haupt, G. Raz, S. Wright, and R. Nowak, "Toeplitz-structured compressed sensing matrices," *14th IEEE/SP Workshop on Statistical Signal Processing, pp. 294-298*, 2007.

[25] W. Bajwa, R. Calderbank, and S. Jafarpour, "Why Gabor Frames? Two Fundamental Measures of Coherence and Their Role in Model Selection," *Journal of Communications and Networking, Vol. 12 (4), pp. 289-307*, 2010.

[26] J. Tropp, "The Sparsity Gap: Uncertainty Principles Proportional to Dimension," *To appear, Proc. 44th Ann. IEEE Conf. Information Sciences and Systems (CISS)*, 2010.

[27] R. Calderbank, S. Howard, and S. Jafarpour, "Sparse Reconstruction via the Reed-Muller Sieve," *Proceedings of IEEE Symposium on Information Theory (ISIT)*, 2010.

[28] A. Kerdock, "A class of low-rate nonlinear binary codes," *Information and Control, Vol. 20, pp.182-187*, 1972.

[29] P. Delsarte and J. M. Goethals, "Alternating bilinear forms over GF(q)," *Journal of Combinatorial Theory, Vol. 19, pp. 26-50*, 1975.

[30] A. R. Hammons, P. V. Kumar, A. R. Calderbank, N. J. A. Sloane, and P. Sole, "The $\mathbb{Z}_4$-linearity of Kerdock Codes, Preparata, Goethals, and related codes," *IEEE Transactions on Information Theory, Vol. 40 (2), pp. 301-319*, March 1994.

[31] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*. North-Holland: Amsterdam, 1977.

[32] J. Kent, "Sparse reconstruction with delsarte-goethals frames and sieves," *Senior Thesis, Department of Mathematics, Princeton University*, May 2010.

[33] S. Howard, R. Calderbank, and S. Searle, "A fast reconstruction algorithm for deterministic compressive sensing using second order Reed-Muller codes," *Conference on Information Sciences and Systems (CISS), Princeton, ISBN: 978-1-4244-2246-3, pp: 11 - 15*, March 2008.

[34] R. Berinde, A. Gilbert, P. Indyk, H. Karloff, and M. Strauss, "Combining geometry and combinatorics: a unified approach to sparse signal recovery.," *46th Annual Allerton Conference on Communication, Control, and Computing, pp. 798-805*, September 2008.

[35] E. Candès and J. Romberg, "ℓ_1-magic: Recovery of sparse signals via convex programming," *available at http://www.acm.caltech.edu/l1magic*, 2005.

[36] S. Wright, R. Nowak, and M. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing, Vol. 57 (7), pp. 2479-2493*, July 2009.

[37] V. Guruswami, J. Lee, and A. Razborov, "Almost euclidean subspaces of ℓ_1 via expander codes," *Proceedings of the 19th annual ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 353-362*, January 2008.

[38] I. Johnstone, "Chi-square oracle inequalities," *In State of the Art in Probability and Statistic, Vol. 37, pp. 399-418*, 2001.

[39] C. McDiarmid, "On the method of bounded differences," *Surveys in combinatorics, pp. 148-188, Cambridge Univ. Press, Cambridge*, 1989.

[40] S. Kutin, "Extensions to McDiarmid's inequality when differences are bounded with high probability," *Technical Report TR-2002-045, University of Chicago.*, April, 2002.

Appendix A
Tail Bounds and Concentration Inequalities

In this appendix, we provide the main concentration inequalities which are used throughout the paper.

Proposition 32 (Gaussian tail bound)**.** *Let $X \approx \mathcal{N}(0, \sigma^2)$ be a zero-mean Gaussian random variable with variance σ^2 Then for all $0 \leq \epsilon$, we have*

$$\Pr\left[|X| \geq \epsilon\sigma\right] \leq 2\exp\left\{-\frac{\epsilon^2}{2}\right\}.$$

Proposition 33 (χ^2-concentration [38])**.** *Let $X \approx \chi^2_m$ be a chi-squared random variable with m degrees of freedom, with mean $m\sigma^2$, and with standard deviation $\sqrt{2m}\sigma^2$. Then for all $0 \leq \epsilon \leq \frac{1}{2}$, we have*

$$\Pr\left[X - m\sigma^2 \geq \epsilon m\sigma^2\right] \leq \exp\left\{-\frac{3}{16}m\epsilon^2\right\}.$$

Proposition 34 (Azuma's Inequality [39])**.** *Suppose $\langle Z_0, Z_1, \cdots, Z_k\rangle$ is a bounded-difference martingale sequence, that is for each i, $\mathbb{E}[Z_i] = Z_{i-1}$, and $|Z_i - Z_{i-1}| \leq c_i$. Then for all $\epsilon > 0$,*

$$\Pr\left[|Z_k - Z_0| \geq \epsilon\right] \leq 2\exp\left\{\frac{-\epsilon^2}{2\sum_{i=1}^k c_i^2}\right\}.$$

In this paper, we use the Azuma's Inequality for complex martingale random variables.

Lemma 35 (Complex Azuma's Inequality)**.** *Let $\langle Z_0, Z_1, \cdots, Z_k\rangle$ be a set of complex random variables such that, for each i, $\mathbb{E}[Z_i] = Z_{i-1}$, and $|Z_i - Z_{i-1}| \leq c_i$. Then for all $\epsilon > 0$,*

$$\Pr\left[|Z_k - Z_0| \geq \epsilon\right] \leq 4\exp\left\{\frac{-\epsilon^2}{8\sum_{i=1}^k c_i^2}\right\}.$$

Proof: For each random variable Z_i let $X_i \doteq \mathrm{Re}(Z_i)$ and $Y_i \doteq \mathrm{Im}(Z_i)$, so that $Z_i = X_i + iY_i$. Then $\mathbb{E}[X_i] = X_{i-1}$ and $\mathbb{E}[Y_i] = Y_{i-1}$. Moreover, by triangle inequality $|X_i - X_{i-1}| \leq |Z_i - Z_{i-1}| \leq c_i$, and $|Y_i - Y_{i-1}| \leq |Z_i - Z_{i-1}| \leq c_i$. Hence, $\langle X_0, \cdots, X_m\rangle$, and $\langle Y_0, \cdots, Y_m\rangle$ form martingale sequences. Now from the triangle inequality we have

$$\Pr\left[|Z_k - Z_0| \geq \epsilon\right] \leq \Pr\left[|X_m - X_0| \geq \frac{\epsilon}{2}\right] + \Pr\left[|Y_m - Y_0| \geq \frac{\epsilon}{2}\right] \leq 4\exp\left\{\frac{-\epsilon^2}{8\sum_{i=1}^k c_i^2}\right\}.$$

■

Proposition 36 (Extension to Azuma's inequality when differences are bounded with high probability [40])**.** *Suppose $\langle Z_0, Z_1, \cdots, Z_k\rangle$ is a martingale sequence, that is for each i, $\mathbb{E}[Z_i] = Z_{i-1}$. Moreover, suppose that with probability $1 - \delta$ for all i: $|Z_i - Z_{i-1}| \leq c_i$, and always $|Z_i - Z_{i-1}| \leq b_i$. Then for every $\epsilon > 0$,*

$$\Pr\left[|Z_k - Z_0| \geq \epsilon\right] \leq 2\left(\exp\left\{\frac{-\epsilon^2}{8\sum_{i=1}^k c_i^2}\right\} + \delta\sum_{i=1}^k \frac{b_i}{c_i}\right).$$

Both the original Azuma inequality and its extension given in Proposition 36 can be applied separately to the real and imaginary parts of a complex random variable to prove concentration about the expected value.

Lemma 37. *Suppose $\langle Z_0, Z_1, \cdots, Z_k\rangle$ is a complex martingale sequence, that is for each i, $\mathbb{E}[Z_i] = Z_{i-1}$. Moreover, suppose that with probability $1-\delta$ for all i: $|Z_i - Z_{i-1}| \leq c_i$, and always $|Z_i - Z_{i-1}| \leq b_i$. Then for all $\epsilon > 0$,*

$$\Pr\left[|Z_k - Z_0| \geq \epsilon\right] \leq 4\left(\exp\left\{\frac{-\epsilon^2}{32\sum_{i=1}^k c_i^2}\right\} + \delta\sum_{i=1}^k \frac{b_i}{c_i}\right).$$

Appendix B
Proof of the Theorem 16

The following lemma is central to many arguments throughout the paper.

Lemma 38. *Let ϵ be a positive number, and let $0 \leq o \leq k-1$. Let $h : [\mathcal{C}]^o \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ be a (η, γ, N)-StRIP-able function. Let $k \leq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, \epsilon^2 N^{2\gamma}\right\}$. Fix the values $\alpha_1, \cdots, \alpha_k$, and fix $o + 1 \leq w \leq \mathcal{C}$. Then*

$$\Pr_{\pi}\left[\left|\sum_{\substack{j=1\\ j\neq w}}^k \alpha_j h(\pi_{1\to o}, \pi_w, \pi_j)\right| \geq 2\epsilon\|\alpha\|_2\right] \leq 4\,\exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\}.$$

Proof: We condition on the value of $\pi_{1\to o}$. Fix $\pi_{1\to o}$ and let $\pi_{-1\to o}$ denote $\pi - \pi_{1\to o}$.

First, we bound the expectation of the sum:

$$\begin{aligned}
&\mathbb{E}_{\pi_{-1\to o}}\left|\sum_{\substack{j=1\\ j\neq w}}^k \alpha_j h(\pi_{1\to o}, \pi_w, \pi_j)\right| \\
&= \mathbb{E}_{\pi_{-1\to o}}\left|\sum_{j=1}^k \alpha_j h(\pi_{1\to o}, \pi_w, \pi_j)(1 - \delta_{w,j})\right| \\
&= \left|\sum_{j=1}^k \alpha_j (1 - \delta_{w,j})\mathbb{E}_{\pi_{-1\to o}}\left[h(\pi_{1\to o}, \pi_w, \pi_j)\right]\right| \qquad (14) \\
&\leq \sum_{j=1}^k |\alpha_j|\,(1 - \delta_{w,j})\left|\mathbb{E}_{\pi_{-1\to o}}\left[h(\pi_{1\to o}, \pi_w, \pi_j)\right]\right| \\
&\leq \sum_{j=1}^k |\alpha_j|\,N^{-\gamma} \leq \sqrt{k}N^{-\gamma}\|\alpha\| \leq \epsilon\|\alpha\|
\end{aligned}$$

where in (14) we use linearity of expectation, and the last line follows from the assumption on average coherence, the Cauchy-Schwarz inequality, and the assumption $k \leq \epsilon^2 N^{2\gamma}$.

Next we bound the difference between the sum and its expectation. For this we condition on the value of π_w. Fix π_w (and recall that we have already fixed $\pi_{1\to o}$). For $o+1 \leq t \leq k$

define the martingale sequence

$$Z_t\left(\pi_{o+1\to t-1},\pi_t\right) \doteq \mathbb{E}_{(\pi_{t+1\to k})}\left[\sum_{\substack{j=1\\ j\neq w}}^{k}\alpha_j h(\pi_{1\to o},\pi_w,\pi_j)|\pi_{1\to t}\right],$$

and define

$$Z_o = \mathbb{E}_{\pi_{-1\to o}}\left|\sum_{\substack{j=1\\ j\neq w}}^{k}\alpha_j h(\pi_{1\to o},\pi_w,\pi_j)\right|.$$

Our goal is to bound the difference

$$Z_t\left(\pi_{1\to t-1},\pi_t\right) - Z_{t-1}\left(\pi_{1\to t-1}\right),$$

for all fixed and distinct values $\pi_{1\to t-1},\pi_t$. It follows from the linearity of expectation and marginalization of the probability that

$$\begin{aligned} c_t \doteq |Z_t - Z_{t-1}| &\leq 2|\alpha_t|\,N^{-\eta} + \frac{2(k-t)N^{-\eta}}{\mathcal{C}-t}\|\alpha\|_1 \qquad (15)\\ &\leq 2|\alpha_t|\,N^{-\eta} + \frac{2kN^{-\eta}}{\mathcal{C}-k}\|\alpha\|_1. \end{aligned}$$

In order to use Azuma's inequality, we need to bound $\sum_{t=o}^{k} c_t^2$:

$$\begin{aligned} \sum_{t=o}^{k} c_t^2 &= 4N^{-2\eta}\sum_{t=1}^{k}\left(|\alpha_t| + \frac{k\|\alpha\|_1}{\mathcal{C}-k}\right)^2 \qquad (16)\\ &\leq 4N^{-2\eta}\left(\|\alpha\|_2^2 + \frac{k^4}{(\mathcal{C}-k)^2}\|\alpha\|_2^2 + 2\frac{k^2}{\mathcal{C}-k}\|\alpha\|_2^2\right). \end{aligned}$$

Consequently, if $k \leq \sqrt{\mathcal{C}} - \frac{1}{2}$ then $\sum_{t=o}^{k} c_t^2 \leq 16N^{-2\eta}\|\alpha\|_2^2$. Now it follows from Azuma's inequality (Lemma 35), that for every $\epsilon \geq \sqrt{k}\,N^{-\gamma}$,

$$\begin{aligned} &\Pr_{\pi_{-1\to o}}\left[\left|\sum_{\substack{j=1\\ j\neq w}}^{k}\alpha_j h(\pi_{1\to o},\pi_w,\pi_j)\right| \geq 2\epsilon\|\alpha\|_2\right]\\ \leq\quad & 4\exp\left\{-\frac{N^{2\eta}\epsilon^2\|\alpha\|^2}{128\,\|\alpha\|^2}\right\} \leq 4\exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\} \end{aligned}$$

Taking the expectation of this probability over all values of $\pi_{1\to o}$ completes the proof. ■

Note that the bound on the sum is $2\epsilon\|\alpha\|_2$ and not $\epsilon\|\alpha\|_2$. This is because we use the triangle inequality to combine the concentration of the expectation about zero, and the concentration of the sum about the expectation. If the expectation of the sum is zero, then we only have to use the concentration of the sum about its expectation, and the triangle inequality is unnecessary. Thus, if $\gamma = \infty$ then the sum has a bound of $\epsilon\|\alpha\|_2$ with the same probability.

We now prove Theorems 16 and 17.

Proof of Theorem 16: Suppose there exists an index i, $1 \leq i \leq k$, or an index $o \in [\mathcal{C}] - \pi_{1\to k}$ such that

$$\left|\sum_{j:j\neq i}\alpha_j h(\pi_i,\pi_j)\right| > 2\epsilon\|\alpha\|_2 \text{ or } \left|\sum_{j=1}^{k}\alpha_j h(o,\pi_j)\right| > 2\epsilon\|\alpha\|_2. \qquad (17)$$

In either case there is an index $w \in [\mathcal{C}]$ such that

$$\left|\sum_{\substack{j=1\\ j\neq w}}^{k}\alpha_j h(\pi_w,\pi_j)\right| > 2\epsilon\|\alpha\|_2. \qquad (18)$$

We may use h to define a "virtual" StRIP-able function with three arguments. Taking the union bound over all possible $w \in [\mathcal{C}]$, Lemma 38 implies:

$$\begin{aligned} &\Pr_{\pi}\left[\exists w \in [\mathcal{C}] : \left|\sum_{\substack{j=1\\ j\neq w}}^{k}\alpha_j h(\pi_w,\pi_j)\right| > 2\epsilon\|\alpha\|_2\right] \qquad (19)\\ &\leq 4\mathcal{C}\exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\}. \end{aligned}$$

■

The proof of Theorem 17 is similar:

Proof of Theorem 17: The StRIP condition is broken only if there exists t, $1 \leq t \leq k$ such that

$$\left|\sum_{\substack{j=1\\ j\neq t}}^{k}\alpha_j h(\pi_{1\to t-1},\pi_t,\pi_j)\right| > \epsilon\|\alpha\|_2.$$

Taking the union bound over all t and applying Lemma 38 yields

$$\begin{aligned} &\Pr_{\pi}\left[\exists\, t\ :\left|\sum_{\substack{j=1\\ j\neq t}}^{k}\alpha_j h(\pi_{1\to t-1},\pi_t,\pi_j)\right| > \epsilon\|\alpha\|_2\right]\\ &\leq 4k\exp\left\{-\frac{N^{2\eta}\epsilon^2}{128}\right\}. \end{aligned}$$

■

Appendix C
Proof of Theorem 19

This proof uses a martingale argument similar to that of Lemma 38, and invokes the Extended Azuma's Inequality. We start by defining

$$\tau(\pi_1,\ldots,\pi_k) = \sum_i\sum_{j\neq i}\alpha_i\overline{\alpha_j}h(\pi_i,\pi_j)$$

which is the sum we want to bound. We next define the Martingale sequence

$$Z_t = \mathbb{E}_\pi\left[\tau(\pi_1,\ldots,\pi_k)|\pi_1,\ldots,\pi_k\right]$$

Finally we define four families of functions $\{h_{i,t}\}_{t=1}^{k}$. In the following definitions, and throughout the proof, p, q denote random variables drawn uniformly at random from $[\mathcal{C}]$.

1) For $1 \leq t \leq k$, define $h_{1,t} : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ by

$$h_{1,t}(\pi_{1\to t-1}, \pi_t, \pi_i) := h(\pi_i, \pi_t) - \mathbb{E}_{p\notin\{\pi_{1\to t-1}\}}[h(\pi_i, p)].$$

2) For $1 \leq t \leq k$, define $h_{2,t} : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ by

$$\begin{aligned} &h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \\ &\quad\doteq \mathbb{E}_{p\notin\{\pi_{1\to t}\}}[h(\pi_t, p)] - \mathbb{E}_{\substack{p,q\notin\{\pi_{1\to t}\}\\ p\neq q}}[h(p, q)]. \end{aligned}$$

3) For $1 \leq t \leq k$, define $h_{3,t} : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ by

$$\begin{aligned} &h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \\ &\quad\doteq \mathbb{E}_{p\notin\{\pi_{1\to t}\}}[h(\pi_i, p)] - \mathbb{E}_{p\notin\{\pi_{1\to t-1}\}}[h(\pi_i, p)]. \end{aligned}$$

4) For $1 \leq t \leq k$, define $h_{4,t} : [\mathcal{C}]^{t-1} \times [\mathcal{C}] \times [\mathcal{C}] \to \mathbb{C}$ by

$$\begin{aligned} &h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \\ &\quad\doteq \mathbb{E}_{\substack{p,q\notin\{\pi_{1\to t}\}\\ p\neq q}}[h(p, q)] - \mathbb{E}_{\substack{p,q\notin\{\pi_{1\to t-1}\}\\ p\neq q}}[h(p, q)]. \end{aligned}$$

Note that we consider $h_{2,t}$ and $h_{4,t}$ as functions of π_i, although they do not truly depend on π_i.

Lemma 39. *For* $1 \leq t \leq k$

$$\begin{aligned} &|Z_t - Z_{t-1}| \\ &\leq 2|\alpha_t| \left| \sum_{i=1}^{t-1} \alpha_i h_{1,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\ &+ 2|\alpha_t| \left| \sum_{i=t+1}^{k} \alpha_i h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\ &+ 2\|\alpha\|_1 \left| \sum_{i=1}^{t-1} \alpha_i h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\ &+ \|\alpha\|_1 \left| \sum_{i=t+1}^{k} \alpha_i h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \end{aligned}$$

Proof: Using the linearity of expectation and the assumption of a uniform distribution for π we can write

$$\begin{aligned} Z_t &= \sum_{i=1}^{t} \sum_{\substack{j=1\\ j\neq i}}^{t} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j) \\ &+ \sum_{i=1}^{t} \sum_{j=t+1}^{k} \alpha_i \overline{\alpha_j} \mathbb{E}_{p\notin\{\pi_1,\dots,\pi_t\}}[h(\pi_i, p)] \\ &+ \sum_{i=t+1}^{k} \sum_{j=1}^{t} \alpha_i \overline{\alpha_j} \mathbb{E}_{p\notin\{\pi_1,\dots,\pi_t\}}[h(p, \pi_j)] \\ &+ \sum_{i=t+1}^{k} \sum_{\substack{j=t+1\\ j\neq i}}^{k} \alpha_i \overline{\alpha_j} \mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_t\}\\ p\neq q}}[h(p, q)] \qquad (20) \end{aligned}$$

where p, q are uniformly distributed over $[\mathcal{C}]$. Applying the skew-symmetry of h and using the fact that $z + \overline{z} = 2\Re(z)$, we can rewrite this as

$$\begin{aligned} Z_t &= \sum_{i=1}^{t} \sum_{\substack{j=1\\ j\neq i}}^{t} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j) \\ &+ 2\Re\left(\left(\sum_{j=t+1}^{k} \overline{\alpha_j} \right) \left(\sum_{i=1}^{t} \alpha_i \mathbb{E}_{p\notin\{\pi_1,\dots,\pi_t\}}[h(\pi_i, p)] \right) \right) \\ &+ \left(\sum_{i=t+1}^{k} \sum_{\substack{j=t+1\\ j\neq i}}^{k} \alpha_i \overline{\alpha_j} \right) \left(\mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_t\}\\ p\neq q}}[h(p, q)] \right). \qquad (21) \end{aligned}$$

Next we can find the difference $Z_t - Z_{t-1}$ by examining the difference of each of the three terms in 21 with their corresponding versions when t is set to $t - 1$. The first term becomes

$$\begin{aligned} &\sum_{i=1}^{t} \sum_{\substack{j=1\\ j\neq i}}^{t} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j) - \sum_{i=1}^{t-1} \sum_{\substack{j=1\\ j\neq i}}^{t-1} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j) \\ &= \sum_{i=1}^{t-1} \alpha_i \overline{\alpha_t} h(\pi_i, \pi_t) + \sum_{j=1}^{t-1} \alpha_t \overline{\alpha_j} h(\pi_t, \pi_j) \\ &= 2\Re\left(\overline{\alpha_t} \sum_{i=1}^{t-1} \alpha_i h(\pi_i, \pi_t) \right) \qquad (22) \end{aligned}$$

The second term becomes

$$\begin{aligned} &2\Re\left(\left(\sum_{j=t+1}^{k} \overline{\alpha_j} \right) \left(\sum_{i=1}^{t} \alpha_i \mathbb{E}_{p\notin\{\pi_1,\dots,\pi_t\}}[h(\pi_i, p)] \right) \right) \\ &- 2\Re\left(\left(\sum_{j=t}^{k} \overline{\alpha_j} \right) \left(\sum_{i=1}^{t-1} \alpha_i \mathbb{E}_{p\notin\{\pi_1,\dots,\pi_{t-1}\}}[h(\pi_i, p)] \right) \right) \\ &\qquad (23) \end{aligned}$$

The third term becomes

$$\begin{aligned} &\left(\sum_{i=t+1}^{k} \sum_{\substack{j=t+1\\ j\neq i}}^{k} \alpha_i \overline{\alpha_j} \right) \left(\mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_t\}\\ p\neq q}}[h(p, q)] \right) \\ &- \left(\sum_{i=t}^{k} \sum_{\substack{j=t\\ j\neq i}}^{k} \alpha_i \overline{\alpha_j} \right) \left(\mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_{t-1}\}\\ p\neq q}}[h(p, q)] \right) \\ &= \left(\mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_t\}\\ p\neq q}}[h(p, q)] - \mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_{t-1}\}\\ p\neq q}}[h(p, q)] \right) \\ &\left(\sum_{i=t+1}^{k} \sum_{\substack{j=t+1\\ j\neq i}}^{k} \alpha_i \overline{\alpha_j} \right) \\ &- 2\Re\left(\mathbb{E}_{\substack{p,q\notin\{\pi_1,\dots,\pi_{t-1}\}\\ p\neq q}}[h(p, q)] \left(\alpha_t \sum_{j=t+1}^{k} \overline{\alpha_j} \right) \right) \qquad (24) \end{aligned}$$

We can now use the triangle inequality to bound $|Z_t - Z_{t-1}|$:

$$\begin{aligned}
&|Z_t - Z_{t-1}| \\
&\leq 2|\alpha_t| \left| \sum_{i=1}^{t-1} \alpha_i h_{1,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\
&+ 2|\alpha_t| \left| \sum_{i=t+1}^{k} \alpha_i h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\
&+ 2\|\alpha\|_1 \left| \sum_{i=1}^{t-1} \alpha_i h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \\
&+ \|\alpha\|_1 \left| \sum_{i=t+1}^{k} \alpha_i h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i) \right| \qquad (25)
\end{aligned}$$

■

Lemma 40. *For $1 \leq t \leq k$ the following StRIP-ability conditions are satisfied*

1) *The function $\frac{1}{2}h_{1,t}$ is (η,∞)-StRIP-able.*
2) *The function $\frac{1}{2}h_{2,t}$ is (η,∞)-StRIP-able.*
3) *The function $\frac{k^2}{2}h_{3,t}$ is (η,∞)-StRIP-able.*
4) *The function $\frac{k^2}{4}h_{4,t}$ is (η,∞)-StRIP-able.*

Proof: By linearity of expectation, for any fixed $\{\pi_1,\ldots,\pi_{t-1}\}$ we can write

$$\mathbb{E}_{\pi_t \notin \{\pi_1,\ldots,\pi_{t-1}\}}[h_{1,t}(\pi_{1\to t-1}, \pi_t, \pi_i)] = 0$$

Also, by worst-case coherence of h, $|h_{1,t}(\pi_{1\to t-1}, \pi_t, \pi_i)| \leq 2N^{-\eta}$, so $\frac{1}{2}h_{1,t}$ is (η,∞)-StRIP-able.

Next, $h_{2,t}$ can be rewritten as

$$\begin{aligned}
h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i) &= \mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_t\}}[h(\pi_t, p)] \\
&- \mathbb{E}_{\pi_t\notin\{\pi_1,\ldots\pi_{t-1}\}}\left[\mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_t\}}[h(\pi_t,p)]\right]
\end{aligned}$$

so for any fixed $\{\pi_1,\ldots,\pi_{t-1}\}$ we have

$$\mathbb{E}_{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\}} h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i) = 0.$$

Again using worst-case coherence of h, we find $|h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i)| \leq 2N^{-\eta}$. Thus $\frac{1}{2}h_{2,t}$ is (η,∞)-StRIP-able.

Next we expand $h_{3,t}$ for $t \neq i$:

$$\begin{aligned}
&|h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i)| \\
=\ & \Pr_{p\notin\pi_{1\to t-1}}[p = \pi_t] \left|\mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_t\}}[h(\pi_i,p)] - h(\pi_i,\pi_t)\right| \\
=\ & \frac{2N^{-\eta}}{\mathcal{C}-(t-1)} \leq \frac{2N^{-\eta}}{\mathcal{C}-k} \leq \frac{2N^{-\eta}}{k^2}
\end{aligned}$$

where in the last step we used the assumption $k \leq \sqrt{\mathcal{C}} - \frac{1}{2}$. We bound average coherence of $h_{3,t}$ by fixing $\{\pi_1,\ldots,\pi_{t-1}\}$:

$$\begin{aligned}
& \mathbb{E}_{\substack{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ \pi_i\neq\pi_t}}[h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i)] \\
=\ & \mathbb{E}_{\substack{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ \pi_i\neq\pi_t}}\left[\mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_t\}}[h(\pi_i,p)]\right] \\
-\ & \mathbb{E}_{\substack{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ \pi_i\neq\pi_t}}\left[\mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_{t-1}\}}[h(\pi_i,p)]\right] \\
=\ & \mathbb{E}_{p\notin\{\pi_1,\ldots,\pi_{t-1}\}}\left[\mathbb{E}_{\substack{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1},p\} \\ \pi_i\neq\pi_t}}[h(\pi_i,p)] - h(\pi_i,p)\right] \\
=\ & 0
\end{aligned}$$

Hence $\frac{k^2}{2}h_{3,t}$ is (η,∞)-StRIP-able.

Next, for any fixed $\{\pi_1,\ldots,\pi_{t-1}\}$,

$$\begin{aligned}
&\mathbb{E}_{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\}}[h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i)] \\
&= \mathbb{E}_{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\}}\left[\mathbb{E}_{\substack{p,q\notin\{\pi_1,\ldots,\pi_t\} \\ p\neq q}}[h(p,q)]\right] \\
&- \mathbb{E}_{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1}\}}\left[\mathbb{E}_{\substack{p,q\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ p\neq q}}[h(p,q)]\right] \\
&= \mathbb{E}_{\substack{p,q\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ p\neq q}}\left[\mathbb{E}_{\pi_t\notin\{\pi_1,\ldots,\pi_{t-1},p,q\}}[h(p,q)] - h(p,q)\right] \\
&= 0. \qquad (26)
\end{aligned}$$

By conditioning on whether p or q are equal to π_t we can write

$$\begin{aligned}
&|h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i)| \\
&\leq \left(1 - Pr_{\substack{p,q\notin\{\pi_1,\ldots,\pi_{t-1}\} \\ p\neq q}}[p,q\neq\pi_t]\right) \left|\mathbb{E}_{\substack{p,q\notin\{\pi_1,\ldots,\pi_t\} \\ p\neq q}}[h(p,q)]\right| \\
&+ \left|\mathbb{E}_{q\notin\{\pi_1,\ldots,\pi_t\}}[h(\pi_t,q)]\right| \\
&\leq \frac{4N^{-\eta}}{\mathcal{C}-(t-1)} \leq \frac{4N^{-\eta}}{k^2}\frac{k^2}{\mathcal{C}-k} \leq \frac{4N^{-\eta}}{k^2} \qquad (27)
\end{aligned}$$

Hence $\frac{k^2}{4}h_{4,t}$ is (η,∞)-StRIP-able. ■

Theorem 41. *Let h be an (η,γ)-StRIP-able function. Assuming a uniform support model, let α be a k-sparse signal, and fix the values $\alpha_1,\ldots,\alpha_k$. Let $\xi > 0$ such that $k \leq \min\left\{\sqrt{\mathcal{C}} - \frac{1}{2}, \xi N^\gamma\right\}$ Then for any $\epsilon > 0$*

$$\begin{aligned}
&\Pr\left[\left|\sum_i \sum_{j\neq i} \alpha_i \overline{\alpha_j} h(\pi_i,\pi_j)\right| \geq 2\xi\|\alpha\|^2\right] \\
&\leq 4\exp\left\{\frac{-\xi^2}{8192\epsilon^2}\right\} + \frac{56N^{-\eta}k^{\frac{7}{2}}}{\epsilon}\exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\}
\end{aligned}$$

Proof: Combining Lemma 40 and Theorem 17 yields the following probabilistic bounds for all $\epsilon > 0$:

1)

$$\begin{aligned}
&\Pr\left[\exists t : \left|\sum_{i=1}^{t-1} \alpha_i h_{1,t}(\pi_i,\pi_t)\right| \geq 2\epsilon\|\alpha\|_2\right] \\
&\leq 4k\exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\} \qquad (28)
\end{aligned}$$

2)

$$\Pr\left[\exists t : \left|\sum_{i=t+1}^{k} \alpha_i h_{2,t}(\pi_{1\to t-1}, \pi_t, \pi_i)\right| \geq 2\epsilon\|\alpha\|_2\right]$$
$$\leq 4k \exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\} \tag{29}$$

3)

$$\Pr\left[\exists t : \left|\sum_{i=1}^{t-1} \alpha_i h_{3,t}(\pi_{1\to t-1}, \pi_t, \pi_i)\right| \geq \frac{2}{k^2}\epsilon\|\alpha\|\right]$$
$$\leq 4k \exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\} \tag{30}$$

4)

$$\Pr\left[\exists t : \left|\sum_{i=t+1}^{k} \alpha_i h_{4,t}(\pi_{1\to t-1}, \pi_t, \pi_i)\right| \geq \frac{4}{k^2}\epsilon\|\alpha\|\right]$$
$$\leq 4k \exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\} \tag{31}$$

Now, taking the union bound over 28, 29, 30 and 31, and applying Lemma 39, we find that with probability $1 - \delta$

$$\begin{aligned} |Z_t - Z_{t-1}| &\leq \epsilon\|\alpha\|\left(4|\alpha_t| + 4|\alpha_t| + \frac{4}{k^2}\|\alpha\|_1 + \frac{4}{k^2}\|\alpha\|_1\right) \\ &= 8\epsilon\|\alpha\|\left(|\alpha_t| + \frac{\|\alpha\|_1}{k^2}\right) \end{aligned} \tag{32}$$

where

$$\delta \leq 16k \exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\} \tag{33}$$

Referring back to Lemma 39, we also have the following absolute bound on the martingale sequence:

$$|Z_t - Z_{t-1}| \leq \sqrt{k}N^{-\eta}\|\alpha\|\left(4|\alpha_t| + 3\|\alpha\|_1\right) \tag{34}$$

By linearity of expectation, we can bound the expectation of τ:

$$\mathbb{E}_\pi\left|\sum_i \sum_{j\neq i} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j)\right| \leq kN^{-\gamma}\|\alpha\|^2$$

Therefore, for any $\xi \geq kN^{-\gamma}$, we can use 32, 33 and 34 with the Extended Azuma's Inequality:

$$\Pr\left[\left|\sum_i \sum_{j\neq i} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j)\right| \geq 2\xi\|\alpha\|^2\right]$$
$$\begin{aligned} \leq\ & 4\exp\left\{\frac{-\xi^2\|\alpha\|^4}{32\sum_{t=1}^{k}\left(8\epsilon\|\alpha\|\left(|\alpha_t| + \frac{1}{k^2}\|\alpha\|_1\right)\right)^2}\right\} \\ +\ & 64k\exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\}\sum_{t=1}^{k}\frac{\sqrt{k}N^{-\eta}\|\alpha\|\left(4|\alpha_t| + 3\|\alpha\|_1\right)}{8\epsilon\|\alpha\|\left(|\alpha_t| + \frac{1}{k^2}\|\alpha\|_1\right)} \end{aligned}$$

Now, by application of the Cauchy-Schwarz Inequality,

$$\begin{aligned} &\sum_{t=1}^{k}\left(8\epsilon\|\alpha\|\left(|\alpha_t| + \frac{1}{k^2}\|\alpha\|_1\right)\right)^2 \\ =\ & 64\epsilon^2\|\alpha\|^2\sum_{t=1}^{k}\left(|\alpha_t|^2 + \frac{2}{k^2}|\alpha_t|\|\alpha\|_1 + \frac{1}{k^4}\|\alpha\|_1^2\right) \\ =\ & 64\epsilon^2\|\alpha\|^2\left(\|\alpha\|^2 + \frac{2}{k^2}\|\alpha\|_1^2 + \frac{1}{k^3}\|\alpha\|_1^2\right) \\ \leq\ & 64\epsilon^2\|\alpha\|^2\left(\|\alpha\|^2 + \frac{2}{k}\|\alpha\|^2 + \frac{1}{k^2}\|\alpha\|^2\right) \\ \leq\ & 64\epsilon^2\|\alpha\|^4\left(1 + \frac{1}{k}\right)^2 \\ \leq\ & 256\epsilon^2\|\alpha\|^4 \end{aligned}$$

and

$$\begin{aligned} \frac{\sqrt{k}N^{-\eta}\|\alpha\|\left(4|\alpha_t| + 3\|\alpha\|_1\right)}{8\epsilon\|\alpha\|\left(|\alpha_t| + \frac{1}{k^2}\|\alpha\|_1\right)} &\leq \frac{\sqrt{k}N^{-\eta}}{8\epsilon}\frac{7\|\alpha\|_1}{\frac{1}{k^2}\|\alpha\|_1} \\ &\leq \frac{7k^{\frac{5}{2}}N^{-\eta}}{8\epsilon}. \end{aligned}$$

So we can write

$$P\left[\left|\sum_i \sum_{j\neq i} \alpha_i \overline{\alpha_j} h(\pi_i, \pi_j)\right| \geq 2\xi\|\alpha\|^2\right]$$
$$\leq\ 4\exp\left\{\frac{-\xi^2}{8192\epsilon^2}\right\} + \frac{56N^{-\eta}k^{\frac{7}{2}}}{\epsilon}\exp\left\{\frac{-N^{2\eta}\epsilon^2}{128}\right\}$$

■

APPENDIX D
PROOF OF LEMMA 24

We start by analyzing the first term in (12). It follows from expanding the expectation that

$$\begin{aligned} &\sum_{i=1}^{k}\frac{|\alpha_i|^2}{\sqrt{N}}\left(1 - \delta_{\Delta,\pi_i}\right)\mathbb{E}_a\left[i^{a\left(P_{\pi_i} - P_\Delta\right)a^\top}\delta_{a,\pi_i}^{aP_\Delta}\right] \\ &= \sum_{i=1}^{k}\frac{|\alpha_i|^2}{\sqrt{N^3}}\left(1 - \delta_{\Delta,\pi_i}\right)\sum_{\substack{a \\ aP_\Delta = aP_{\pi_i}}} i^{a\left(P_{\pi_i} - P_\Delta\right)a^\top}. \end{aligned}$$

Since $\Delta \neq \pi_i$, the null space of $P_\Delta + P_{\pi_i}$ has dimension at most $2r$, and there are at most 2^{2r} vectors a for which $aP_\Delta = aP_{\pi_i}$. Hence

$$\left|\left(1 - \delta_{\Delta,\pi_i}\right)\sum_{\substack{a \\ aP_\Delta = aP_{\pi_i}}} i^{a\left(P_{\pi_i} - P_\Delta\right)a^\top}\right| \leq 2^{2r},$$

and

$$\left|\sum_{i=1}^{k}\frac{|\alpha_i|^2}{\sqrt{N}}\left(1 - \delta_{\Delta,\pi_i}\right)\mathbb{E}_a\left[i^{a\left(P_{\pi_i} - P_\Delta\right)a^\top}\delta_{a,\pi_i}^{aP_\Delta}\right]\right|$$
$$\leq \sum_{i=1}^{k}\frac{|\alpha_i|^2}{N^{\frac{3}{2} - \frac{2r}{m}}}.$$

Next we bound the second term

$$\frac{1}{N^{\frac{3}{2}}}\sum_{i=1}^{k}\sum_{j\neq i}\alpha_i\overline{\alpha_j}\sum_{x} i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}\mathcal{E}_x(\pi_i,\Delta). \tag{35}$$

We need to analyze two distinct cases:

Case 1: All $\pi_1\cdots,\pi_k$ are distinct from Δ. In this case we define the function

$$h\left(\pi_i,\pi_j\right)\doteq\sum_{x} i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}\mathcal{E}_x(\pi_i,\Delta).$$

It follows from Lemma 7 that

$$|h(\pi_i,\pi_j)|\le m\sqrt{|\mathcal{N}(P_{\pi_i}-P_\Delta)|\,|\mathcal{N}(P_{\pi_j}-P_\Delta)|}\le m\,2^{2r}. \tag{36}$$

The extra term m in (36) is a consequence of removing $m+1$ rows from the Reed-Muller sieve (see Remark 9). Now we prove the bound for the average coherence. We rewrite $h(\pi_i,\pi_j)$ as

$$\frac{1}{N}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\sum_{a} i^{(a+x)\left(P_{\pi_i}-P_\Delta\right)(a+x)^\top}.$$

Note that as a ranges over the additive group $\mathbb{F}_2^m$, $a+x$ also ranges over $\mathbb{F}_2^m$. Therefore $\sum_a i^{(a+x)\left(P_{\pi_i}-P_\Delta\right)(a+x)^\top}$ is a column sum, independent of the choice of j. By Lemma 6, its magnitude is smaller than $N^{\frac{1}{2}+\frac{2r}{m}}$. As a result

$$\left|\mathbb{E}_{\substack{\pi\\ j\neq i}} h(\pi_i,\pi_j)\right|\le N^{\frac{1}{2}+\frac{2r}{m}}\left|\mathbb{E}_{\substack{\pi\\ j\neq i}}\left[\frac{1}{N}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\right]\right|.$$

Lemma 5 then implies that

$$\left|\mathbb{E}_{\substack{\pi\\ j\neq i}}\left[\frac{1}{N}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\right]\right|\le\frac{1}{\mathcal{C}-1}.$$

Consequently h is $\left(\frac{1}{2}+r\left(1-\frac{2}{m}\right),-\left(\frac{2r+\log m}{m}\right),N\right)$ StRIP-able. Now let ξ be a positive numbers such that $k\le\min\left\{\sqrt{\mathcal{C}}-\frac{1}{2},\xi N^{\frac{1}{2}+r\left(1-\frac{2}{m}\right)}\right\}$. It follows from Theorem 19 that for any $\epsilon>0$, with probability at least

$$1-4\exp\left\{\frac{-\xi^2N^{2\eta}}{8192\epsilon^2}\right\}+\frac{56N^{-\eta}k^{\frac{7}{2}}}{\epsilon}\exp\left\{\frac{-\epsilon^2N^{2\eta}}{128}\right\},$$

we have

$$\left|\left(\sum_{i=1}^{k}\alpha_i\sum_{j=1}^{k}\overline{\alpha}_j h\left(\pi_i,\pi_j\right)\right)-\|\alpha_{1\to k}\|^2\right|\le 2\xi\|\alpha_{1\to k}\|^2.$$

In particular, by setting $\epsilon=O\left(m\,N^{\frac{2r}{m}}\sqrt{\log\mathcal{C}}\right)$, and $\xi=O\left(m\,N^{\frac{2r}{m}}\log\mathcal{C}\right)$, we can guarantee that with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$

$$\frac{1}{N^{\frac{3}{2}}}\left|\sum_{i}\sum_{j\neq i}\alpha_i\overline{\alpha_j}h(\pi_i,\pi_j)\right|\preceq\frac{m\log\mathcal{C}\|\alpha_{1\to k}\|^2}{N^{\frac{3}{2}-\frac{2r}{m}}}.$$

Case 2: There exists an index t such that π_t equals Δ. In this case (35) can be rewritten as

$$\frac{1}{N^{\frac{3}{2}}}\sum_{\substack{i\\ \pi_i\neq\Delta}}\sum_{j\neq i}\alpha_i\overline{\alpha_j}\sum_{x} i^{x\left(P_{\pi_i}-P_{\pi_j}\right)x^\top}\mathcal{E}_x(\pi_i,\Delta) \tag{37}$$

$$+\frac{1}{N^{\frac{3}{2}}}\alpha_t\sum_{j\neq t}\overline{\alpha_j}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}.$$

An argument similar to the argument used in case 1 can be used to bound the first term in (37). Now we bound the second term. Let ϵ' be any positive number such that $k\le\min\left\{\sqrt{\mathcal{C}}-\frac{1}{2},\epsilon'^2N^{2(r+1)}\right\}$. It follows from Lemma 38 that with probability $1-4\,k\,\exp\left\{\frac{\epsilon'^2N^{1-\frac{2r}{m}}}{128}\right\}$,

$$\frac{1}{N^{\frac{3}{2}}}\left|\alpha_t\sum_{j\neq t}\overline{\alpha_j}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\right|\le\frac{2\epsilon'|\alpha_t|\|\alpha\|_2}{N^{\frac{1}{2}}}.$$

As a result, as long as $\epsilon'\le\frac{1}{20}\sqrt{\frac{\mathrm{MAR}}{k}}$, we get the bound

$$\frac{1}{N^{\frac{3}{2}}}\left|\alpha_t\sum_{j\neq t}\overline{\alpha_j}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\right|\le\frac{|\alpha_t|^2}{10N^{\frac{1}{2}}}.$$

Since $\mathrm{MAR}\succeq\frac{k\log\mathcal{C}}{N^{1-\frac{2r}{m}}}$, by selecting $\epsilon'=O\left(\frac{\sqrt{\log\mathcal{C}}}{N^{\frac{1}{2}-\frac{r}{m}}}\right)$ we guarantee that with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$,

$$\frac{1}{N^{\frac{3}{2}}}\left|\alpha_t\sum_{j\neq t}\overline{\alpha_j}\sum_{x} i^{x\left(P_\Delta-P_{\pi_j}\right)x^\top}\right|\le\frac{|\alpha_t|^2}{10N^{\frac{1}{2}}}.$$

APPENDIX E
PROOF OF LEMMA 25

We expand $\mathbb{E}_a\left[\frac{i^{-aP_\Delta a^\top}}{\sqrt{N}}\sum_x y(x+a)\overline{u(x)}(-1)^{aP_\Delta x^\top}\right]$ as

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{k}\alpha_i\left(\sum_{x}\frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta)\right). \tag{38}$$

There are two distinct cases to handle:

Case 1: All $\pi_1\cdots,\pi_k$ are distinct from Δ. Define the function

$$\hbar\left(\pi_i\right)\doteq\sum_{x}\frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta). \tag{39}$$

First we analyze the average behavior of $\hbar$. We rewrite Equation (39) as

$$\hbar\left(\pi_i\right)\doteq\frac{1}{N^{\frac{3}{2}}}\sum_{x} i^{xP_\Delta x^\top}\overline{u(x)}\sum_{a}\left[i^{(x+a)\left(P_{\pi_i}-P_\Delta\right)(x+a)^\top}\right]$$

Note that as a ranges uniformly over the additive group $\mathbb{F}_2^m$, $a+x$ also ranges uniformly over the group. Hence

$$\begin{aligned}&\left|\mathbb{E}_\pi\left[h(\pi_i)\right]\right|\\ &\leq \frac{\left|\sum_x i^{xP_\Delta x^\top}\overline{u(x)}\right|\,\left|\mathbb{E}_\pi\left[\sum_{y\in\mathbb{F}_2^m} i^{y\left(P_{\pi_i}-P_\Delta\right)y^\top}\right]\right|}{N^{\frac{3}{2}}}\\ &\leq \frac{\|u\|_1}{N^{\frac{1}{2}}}\frac{1}{\mathcal{C}-1}.\end{aligned}$$

Next we provide a uniform upper-bound for the magnitude of $\hbar$. By expanding the expectation in Equation (39) we rewrite $\hbar\left(\pi_i\right)$ as

$$\sum_x \frac{i^{xP_{\pi_i}x^\top}}{N^{\frac{3}{2}}}\overline{u(x)}\sum_a\left(i^{a\left(P_{\pi_i}-P_\Delta\right)a^\top}(-1)^{\left(aP_{\pi_i}-aP_\Delta\right)x^\top}\right). \tag{40}$$

By Proposition 6, the inner sum is either 0 or it has size $2^{\frac{m+t}{2}}$ for some t at most $2r$. The non-zero terms occur at indices x that are elements of a translate of some $m-t$-dimensional subspace. As a result, it follows from the Cauchy-Schwarz inequality that:

$$\max_i |\hbar(\pi_i)| \leq \frac{\|u\|_2}{N^{\frac{3}{2}}}2^{\frac{m+t}{2}}2^{\frac{m-t}{2}} \leq \frac{\|u\|_2}{N^{\frac{1}{2}}}.$$

Having analyzed $\mathbb{E}_{\pi_i}\left[\hbar(\pi)\right]$, and $\max_{\pi_i}\hbar(\pi)$, we can now use Azuma's inequality (Lemma 35) to bound the term $\frac{1}{\sqrt{N}}\sum_{i=1}^k \alpha_i\hbar(\pi_i)$. Let ϵ be any positive number such that $k \leq \min\left\{\frac{\epsilon^2 N^{2(r+2)}}{\|u\|_1^2}, \frac{\mathcal{C}-1}{2}\right\}$. Then

$$\Pr\left[\left|\frac{1}{\sqrt{N}}\sum_{i=1}^k \alpha_i\hbar(\pi_i)\right| \geq 2\epsilon\|\alpha\|\right] \leq 4\exp\left\{\frac{-\epsilon^2N^2}{\|u\|^2 128}\right\}.$$

Now the union bound over all possible choices of Δ implies that with probability at least $1-4\mathcal{C}\exp\left\{\frac{-\epsilon^2N^2}{\|u\|^2128}\right\}$, for every index Δ

$$\frac{1}{\sqrt{N}}\left|\sum_{i=1}^k \alpha_i\left(\sum_x \frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta)\right)\right| \leq 2\epsilon\|\alpha\|.$$

In particular, if $\epsilon = O\left(\frac{\sqrt{\log\mathcal{C}}\|u\|_2}{N}\right)$ then the condition $k \leq \frac{\|u\|_2^2N^{2(r+1)}\log\mathcal{C}}{\|u\|_1^2}$ is always satisfied; consequently, with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, for every index Δ

$$\left|\frac{1}{\sqrt{N}}\sum_{i=1}^k \alpha_i\hbar(\pi_i)\right| \preceq \frac{\sqrt{\log\mathcal{C}}\|u\|_2\|\alpha\|_2}{N}. \tag{41}$$

Now we show that if the elements of u have independent and random sign, then we can use the union bound to strengthen the bound of Equation 41. Let ϵ be a positive number, then using the Gaussian tail bound (Proposition 32)

$$\begin{aligned}\Pr_u\left[\exists\pi_i:\ |\hbar(\pi_i)| \geq \epsilon\|u\|\right] &\leq \sum_{i=1}^{\mathcal{C}}\Pr_u\left[|\hbar(\pi_i)| \geq \epsilon\|u\|\right]\\ &\leq 4\mathcal{C}\exp\left\{\frac{-\epsilon^2N^{2-\frac{2r}{m}}}{16}\right\}.\end{aligned}$$

Now let ξ be any positive number with

$$k \leq \min\left\{\frac{\xi^2N^{2(r+2)}}{\|u\|_1^2}, \frac{\mathcal{C}-1}{2}\right\},$$

with probability at least

$$4\mathcal{C}\left(\exp\left\{\frac{-\xi^2\|\alpha\|^2N}{128\,\epsilon^2\|\alpha\|^2\|u\|^2}\right\} + \mathcal{C}\exp\left\{\frac{-\epsilon^2N^{2-\frac{2r}{m}}}{16}\right\}\right),$$

for every column index Δ we have

$$\frac{1}{\sqrt{N}}\left|\sum_{i=1}^k \alpha_i\left(\sum_x \frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta)\right)\right| \leq 2\xi\|\alpha\|_2.$$

Note that by choosing $\epsilon = O\left(\frac{\sqrt{\log\mathcal{C}}}{N^{1-\frac{r}{m}}}\right)$ and $\xi = O\left(\frac{\log\mathcal{C}\|u\|_2}{N^{\frac{3}{2}-\frac{r}{m}}}\right)$ it is guaranteed that with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, for every index Δ

$$\frac{1}{\sqrt{N}}\left|\sum_{i=1}^k \alpha_i\sum_x \frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta)\right| \preceq \frac{\log\mathcal{C}\|\alpha\|_2\|u\|_2}{N^{\frac{3}{2}-\frac{r}{m}}}.$$

case 2: There exists an index t such that π_t equals Δ. In this case (38) can be written as

$$\begin{aligned}&\frac{1}{\sqrt{N}}\sum_{\substack{i\\ \pi_i\neq\Delta}}\alpha_i\sum_x\frac{i^{xP_{\pi_i}x^\top}}{\sqrt{N}}\overline{u(x)}\mathcal{E}_x(\pi_i,\Delta) \qquad (42)\\ &+\frac{1}{\sqrt{N}}\alpha_t\sum_x\frac{i^{xP_{\pi_t}x^\top}}{\sqrt{N}}\overline{u(x)}.\end{aligned}$$

An argument similar to the one used in proving case 1 can be used to bound the first term in (42). Now we bound the second term. First consider the deterministic noise regime, in which that no information about u is known. In this case if $\|\alpha\|_{\min} \geq 10\|u\|_2$, then if follows from the Cauchy-Schwarz inequality that

$$\left|\frac{1}{\sqrt{N}}\alpha_t\sum_x\frac{i^{xP_{\pi_t}x^\top}}{\sqrt{N}}\overline{u(x)}\right| \leq \frac{|\alpha_t|\|u\|_2}{\sqrt{N}} \leq \frac{|\alpha_t|^2}{10\sqrt{N}}.$$

Next consider the stochastic noise regime. In this case, it follows from Azuma's inequality that with probability $1-O\left(\frac{1}{\mathrm{Poly}(\mathcal{C})}\right)$, $\left|\frac{1}{\sqrt{N}}\alpha_t\sum_x\frac{i^{xP_{\pi_t}x^\top}}{\sqrt{N}}\overline{u(x)}\right| \preceq \frac{\sqrt{\log\mathcal{C}}|\alpha_t|\|u\|_2}{N}$. Therefore, as long as $\|\alpha\|_{\min} \geq \frac{10\sqrt{\log\mathcal{C}}\|u\|_2}{\sqrt{N}}$, we guarantee

$$\left|\frac{1}{\sqrt{N}}\alpha_t\sum_x\frac{i^{xP_{\pi_t}x^\top}}{\sqrt{N}}\overline{u(x)}\right| \preceq \frac{\sqrt{\log\mathcal{C}}|\alpha_t|\|u\|_2}{N} \leq \frac{|\alpha_t|^2}{10\sqrt{N}}.$$