Electronic Colloquium on Computational Complexity, Report No. 60 (2007)

Sparse Random Linear Codes are Locally Decodable and Testable

Tali Kaufman* Madhu Sudan†

July 11, 2007

Abstract

We show that random sparse binary linear codes are locally testable and locally decodable (under any linear encoding) with constant queries (with probability tending to one). By sparse, we mean that the code should have only polynomially many codewords. Our results are the first to show that local decodability and testability can be found in random, *unstructured*, codes. Previously known locally decodable or testable codes were either classical algebraic codes, or new ones constructed very carefully.

We obtain our results by extending the techniques of Kaufman and Litsyn [11] who used the MacWilliams Identities to show that "almost-orthogonal" binary codes are locally testable. Their definition of almost orthogonality expected codewords to disagree in $\frac{n}{2} \pm O(\sqrt{n})$ coordinates in codes of block length n. The only families of codes known to have this property were the dual-BCH codes. We extend their techniques, and simplify them in the process, to include codes of distance at least $\frac{n}{2} - O(n^{1-\gamma})$ for any $\gamma > 0$, provided the number of codewords is $O(n^t)$ for some constant t. Thus our results derive the local testability of linear codes from the classical coding theory parameters, namely the *rate* and the *distance* of the codes.

More significantly, we show that this technique can also be used to prove the "self-correctability" of sparse codes of sufficiently large distance. This allows us to show that random linear codes under linear encoding functions are *locally decodable*. This ought to be surprising in that the definition of a code doesn't specify the encoding function used! Our results effectively say that any linear function of the bits of the codeword can be locally decoded in this case.

*MIT CSAIL. kaufmant@mit.edu.

†MIT CSAIL. madhu@mit.edu. Research supported in part by NSF Award CCR 0514915.

ISSN 1433-8092

1 Introduction

In this paper we study the *local decodability* and *local testability* of *random* codes. Our goal is to see what are general conditions under which error-correcting codes exhibit these properties. We start by defining these concepts.

A (binary) code C is a subset of $\{0,1\}^n$, binary strings of length n. The code is termed an *error-correcting code* with (relative Hamming) distance δ where $\delta = \delta(C) = \min_{x \neq y \in C}\{\delta(x,y)\}$. Here $\delta(x,y)$ denotes the fraction of coordinates $i \in \{1,\ldots,n\}$ for which the ith coordinates of x and y (denoted x_i and y_i) are unequal. Throughout the paper we will be interested in infinite family of codes $\{C_n\}_n$ whose length $n \to \infty$ and whose distance is positive, i.e., for every n, $\delta(C_n) \geq \delta_0 > 0$.

Informally, a code is locally testable if it is possible to test membership of a word v in C probabilistically making few probes into v. Formally, for a vector $v \in \{0,1\}^n$, let $\delta(v,C)$ denote its distance to the set C, i.e., $\min_{x \in C}\{\delta(v,x)\}$. For functions $q : (0,1] \to \mathbb{Z}^+$ and $\epsilon : (0,1] \to (0,1]$, [1] a binary code C is said to be (q,ϵ)-*locally testable* if there exists a probabilistic algorithm T called the *tester* that, given a parameter $\delta > 0$ and oracle access to a vector $v \in \{0,1\}^n$, queries the oracle at most $q(\delta)$ times and accepts $v \in C$ with probability one, and rejects codewords at distance greater than δ with probability at least $\epsilon(\delta)$. C is said to be *strongly q-locally testable* if there exists a constant $\epsilon' > 0$ such that the tester above rejects every $v \notin C$ with probability at least $\epsilon' \cdot \delta(v,C)$ (i.e., $q(\delta)$ is a constant, and $\epsilon(\delta)$ is linear in δ).

A code is informally self-correctable, if it is possible to recover any specified coordinate of a codeword of C given oracle access to a slightly corrupted version of the codeword. Formally, a code C is q-*self-correctable* if there exists constants $\tau > 0$ and $\epsilon < \frac{1}{2}$ and a probabilistic algorithm SC called the corrector such that given oracle access to a word $v \in \{0,1\}^n$ that is τ-close to some codeword $c \in C$, and any index $i \in \{1,\ldots,n\}$, $\mathrm{SC}^v(i)$ computes c_i with probability $1-\epsilon$.

Self-correction is closely related to the more popular notion of locally decodable codes. Local decodability is a property of the encoding function used to encode messages and requires the *message* to be locally decodable from a corrupted codeword. Specifically, an encoding function $E : \{0,1\}^k \to \{0,1\}^n$ is q-locally decodable if there exist exists constants $\tau > 0$ and $\epsilon < \frac{1}{2}$ and a probabilistic decoding algorithm D such that, given oracle access to a word $v \in \{0,1\}^n$ that is τ-close to $E(m)$ for some message $m \in \{0,1\}^k$ and an index $i \in \{1,\ldots,k\}$, $D^v(i)$ computes m_i with probability $1-\epsilon$. Self-correction is a property merely of the set of codewords (i.e., the set $C = \{E(m) | m \in \{0,1\}^k\}$) and while it guarantees decodability of every bit of the codeword it makes no statements about the message bits (since these are not even defined). However it is easy to see that E is q-locally decodable if the code $C_E = \{(m, E(m)) | m\}$ is q-self-correctible. This relation suggests that self-correction is a somewhat stronger property,

Local testability and self-correctability of codes have been examined implicitly every since the seminal work of Blum, Luby and Rubinfeld [6] which shows that Hadamard codes[2] are locally-testable and self-correctable. The general notion of locally testable codes is implicit in [1] and was systematically studied in [9]. Locally decodable codes were implicit in [1] and were formally defined in Katz and Trevisan [10]. Due to the central nature of these notions in works on Probabilistically

[1] In this paper, $\mathbb{Z}^+$ denotes the set of non-negative integers and $(0,1]$ the half-open interval of reals greater than 0 and at most 1.

[2] Strictly speaking [6] only show that "Reed-Muller codes of order 1" are locally testable, and this is a strict subclass of Hadamard codes. The entire class of Hadamard codes is shown to be locally testable by the results of [11].

Checkable Proofs, and Private Information Retrieval Schemes there has been a lot of work on these notions in the past few years. By now, a broad family of codes have been shown to be testable and self-correctable. And recently there has been much activity in constructing very efficient codes that are locally testable (see, for example, [9, 3, 4, 5, 8]) and efficient codes that are locally-decodable [2, 14].

Our interest in these properties is not motivated by the efficiency of the codes, but rather by the generality of analysis techniques. All the properties under consideration (decodability, testability and self-correctability) are known to be quite fragile. For instance, a code C may be testable, but subcodes C' of C (i.e., $C' \subseteq C$) may turn out not to be testable. A code C may be locally decodable for some encoding functions but may not be decodable under other encoding schemes. A code C may be self-correctible, but if one adds just one more bit of redundancy to the encoding (thus $|C'| = |C|$, $C' \subseteq \{0,1\}^{n+1}$ and C' projected to the first n coordinates equals C), then the new code need not be self-correctible. Given this fragility, it is understandable that codes with such properties have been constructed quite carefully in the past, even if the efficiency of the parameters is not the prime concern.

Our Results: In this paper we describe results which are contrary to the descriptions above. We do so by considering a restricted class of codes: specifically linear, sparse codes, chosen at random. Recall that a code C is said to be *linear* if for every $x, y \in C$ it is the case that $x + y$ is also in C, where $x + y$ denotes the sum of x and y viewed as vectors in $\mathbb{F}_2^n$. We show that when a sparse linear code is chosen at random, then it is very likely to be locally decodable and self-correctible.

We derive our results by showing that every "sparse", linear code of "large-distance" is testable and self-correctable. To be precise, say that $C \in \{0,1\}^n$ is $f(n)$-*sparse* if $|C| \leq f(n)$. Our main theorem on local testability shows that every polynomially sparse code of distance $\frac{1}{2} - n^{-\gamma}$ is locally testable.

Theorem 1.1 *For every $t < \infty$ and $\gamma > 0$ there exists a constant k such that the following holds: If $C \subseteq \{0,1\}^n$ is n^t-sparse and $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$, then C is strongly k-locally testable.*

To derive this result, we first focus on a subclass of codes where the distance between codewords is sandwiched tightly between $\frac{1}{2} - n^{-\gamma}$ and $\frac{1}{2} + n^{-\gamma}$. We say that a code $C \subseteq \{0,1\}^n$ is said to be ϵ-*biased* if it satisfies $\frac{1}{2} - \epsilon \leq \delta(x,y) \leq \frac{1}{2} + \epsilon$ for every pair $x \neq y \in C$. (ϵ-bias refers to the same concept as defined in [13], though the description there is different. Our notion of ϵ-bias is also a generalization of the notion of almost orthogonal codes in [11]; almost-orthogonal codes are $O(n^{-1/2})$-biased codes.) For ϵ-biased codes we also derive a complementary "self-correction" result.

Theorem 1.2 *For every $t < \infty$ and $\gamma > 0$ there exists a constant $k = k_{t,\gamma} < \infty$ such that the following holds: If $C \subseteq \{0,1\}^n$ is n^t-sparse and $n^{-\gamma}$-biased, then C is k-self-correctable.*

If a code C has small bias and is sparse, then so is C' consisting of the coordinates of C and a few additional coordinates that are linear functions of the previous coordinates. Using this observation we can use the above result on self-correction to show that sparse, low-bias encodings are locally decodable. (See Corollary 4.6.)

Since a random linear code with n^t-codewords is very likely to be $O(\log n/\sqrt{n})$-biased, we immediately get that sparse random linear codes are locally testable and self-correctible with high probability. (See Theorem 7.2.)

It turns out that our results above do not cover the self-correctibility of even the Hadamard code and/or the dual-BCH code. These codes are not small-biased. Indeed they have the "all-ones" vector as codewords, making their bias as bad as they could possibly be. However we note that these codes are derived from $n^{-\gamma}$-biased codes by "closing them under complements". Specifically for a code C, and vector v, let $C + v$ denote the set $\{x + v | x \in C\}$, and let $C||v$ denote the linear span of C and v, i.e, the set $C \cup (C + v)$. Formally, say that a code C is *complementation-closed* if $C = C'||1^n$ for some code C'. We show the following simple extension of Theorem 1.2 which allows use to extend our testing and self-correction results also for the familiar versions of the Hadamard and dual-BCH codes.

Theorem 1.3 *For every t and $\gamma > 0$, there exists k such that the following hold: Let $C = C'||1^n$, be an n^t-sparse complementation-closed code with distance $\frac{1}{2} - n^{-\gamma}$. Then C is k-locally testable and k-self-correctable.*

Techniques: Our techniques build upon those of Kaufman and Litsyn [11] who analyze the local-testability of dual-BCH codes, which form a special family of codes that satisfy the conditions of Theorem 1.2. Their technique revolves around the analysis of the "weight distribution" of the code C (i.e., the distribution of the weight of a randomly chosen codeword of C) and then using the MacWilliams transform (an extension of the Fourier transform) to analyze a natural tester for the code. They then employ known facts about the weight distribution of the dual-BCH codes to derive their results. When testing a word $v \in \{0,1\}^n$ for membership in C, their analysis relies on the ability to relate the number of codewords of a given weight k in the dual of the code C with the number of codewords in the dual of the code $C||v$. (Recall this is the code consisting of the linear span of C and v.)

Our contributions to this technique is three-fold. First we simplify (and slightly generalize) their analysis in the context of small-biased codes (see Theorem 5.5). We don't use any special properties of the weight distribution other than the most "obvious ones" (the number of codewords is small, and all except the zero codeword have weight very close to $n/2$). Our principal tool here is the Johnson bound which bounds the number of codewords in any ball of small radius. By reducing the proof to such simple elements, we are now able to get to a broader class of codes including randomly chosen codes (for the first time).

Next, we observe that the resulting understanding of the dual of sparse, small-biased, linear codes is so good, that we can pin down the weight distribution of $C^\perp$ almost precisely, to within a $1+o(1)$ multiplicative factor. Making this observation explicit allows us to immediately build and analyze self-correctors for all sparse, small-biased, linear codes. Our analysis shows that the natural local tester for these codes has a roughly "pairwise-independent" distribution of queries and so can be used to recover all coordinates with high probability. Such a result is quite novel in the context of self-correcting in that most previous self-correctors were built for structured codes, whose design already allowed for self-correction algorithms. Here we prove it is an "inevitable" consequence of basic properties (distance and density) of the code.

Third, we strengthen the techniques above to cover the case of codes whose *bias* is not necessarily small. Technically this poses a significant challenge in that we no longer have the ability to pin down the weight distribution of the dual of C precisely and so the previous techniques seem to fail. We overcome the challenge by first considering codes of "moderate" bias (say, 1/3-biased codes) of large distance. In this case we give crude approximations, to within $O(1)$ factors, of the weight distribution of the dual of C. However we show that we can still relate the number of codewords

of in the dual of $C||v$ to those in the dual of C quite precisely and thus deriving an analysis for the natural test.

Our final testing theorem is then derived by viewing every code as a small "extension" of a moderately-biased code. We note that every large-distance code C can be written as $C = C'||v$ where C' has moderate bias. We then show how to derive local tests for C from local tests for C' (see Lemma 6.7), which allows us to get local tests for all sparse codes of large distance.

Conclusions and Future Work: Our work produces the first analysis for local testability and correctability of *random* codes. While much of the analysis of the local test is directly based on the prior work of Kaufman and Litsyn [11], our proofs are more self-contained and simpler, and the results are significantly more general. (The only non-trivial fact we seem to be using are bounds on the roots of the Krawtchouk polynomials. All other facts can be proved by elementary means.) The simplicity of the proof offers hope that by further understanding of the Krawtchouk polynomials, one can prove testing and correcting results for denser classes of codes, and this seems to be an important but challenging open problem. Indeed, right now, even generalizing the result to get self-correctors for general (non-biased) sparse codes of relative distance $\frac{1}{2} - n^{-\gamma}$ is open.

Organization of this paper: Section 2 describes the natural tester and self-correctors for the codes considered in this paper. Section 3 lays out our analysis approach, using the MacWilliams Identities, and states some basic properties about sparse codes. Section 4 analyzes the self-corrector for $n^{-\gamma}$-biased codes. Section 5 analyzes the local test for $n^{-\gamma}$-biased codes. Section 6 presents a strong local tester and self-corrector for complementation closed codes. This section also analyzes the local testability of general (non-biased) sparse codes. Section 7 mentions some implications for random codes, showing that sparse random codes are testable and correctible, while less sparse codes are not so.

2 Preliminaries and Overview of main results

The linearity of a code leads to a natural test and self-correction algorithm for the code. We describe these algorithms below. Most of the rest of the paper is devoted to analyzing these natural algorithms for sparse, low-biased codes (though in Section 6 we discuss some slightly more general codes).

We start with some standard notions and notations: For a positive integer n let $[n]$ denote the set $\{1, \ldots, n\}$. For a word $y \in \{0,1\}^n$ we let the weight of y, denoted $\mathrm{wt}(y)$, be the quantity $|\{i \in [n] \mid y_i = n\}| = \delta(y, 0^n) \cdot n$. For a code $C \subseteq \{0,1\}^n$, let $[C]_k$ denote the codewords of C of weight k, i.e., $[C]_k = \{x \in C \mid \mathrm{wt}(x) = k\}$. For $C \subseteq \{0,1\}^n$, $0 \leq k \leq n$ and $i \in [n]$, let $[C]_{k,i} = \{y \in [C]_k | y_i = 1\}$.

For vectors $x, y \in \{0,1\}^n$ let $\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i (\mathrm{mod} 2)$. For a linear code C, the dual of C is the $C^\perp = \{y \in \{0,1\}^n \mid \langle x, y \rangle = 0 \ \forall x \in C\}$. Straightforward linear-algebra yields that if C is a linear code of dimension ℓ (i.e., $|C| = 2^\ell$), then $C^\perp$ is a linear code of dimension $n - \ell$.

For a finite set S, we let $Z \in_U S$ denote the random variable chosen uniformly from S.

The Self Corrector: We start by describing the canonical k-self-correction algorithm for a binary linear code.

- k-self-corrector $\mathrm{SC}_k^v(i)$
  /* with oracle access to $v \in \{0,1\}^n$ satisfying $\delta(x,v) < \frac{1}{2k}$ for $x \in C$ */.
  - Pick $y \in_U [C^\perp]_{k,i}$.
  - Output $\sum_{\{j \in [n]-\{i\} \mid y_j=1\}} v_j$.

It is straightforward to see that the self-corrector makes $k-1$ queries into v. The harder part of the analysis is to show that the algorithm above correctly recovers x_i for every i, with high probability over its random choices. Indeed it is not clear that the self-corrector does anything. For all we know $[C^\perp]_k$, and hence $[C^\perp]_{k,i}$, may well be empty.

In Section 3, we will prove first that the set if C is sparse and small biased then $[C^\perp]_k$ is indeed non-empty (see Lemma 3.5). In fact, we get very tight estimates on the size of this set. This section introduces notions and techniques needed in the rest of the paper as well.

We then go on to analyze the self-corrector in Section 4. The crucial issue here is to prove that for every i, the set $[C^\perp]_{k,i}$ is non-empty and that the set of locations probed by the algorithm above is roughly uniform. Since the code C, and hence $C^\perp$ are not really under our control, this has to be proven for every code of the given parameters.

The Local Tester: We now describe the *canonical* k-local test T_k associated with a linear code C.

- k-local tester $T_k^v()$ /* with oracle access to $v \in \{0,1\}^n$.
  - Pick $y \in_U [C^\perp]_k$.
  - Accept if and only if $\langle y, v \rangle = 0$.

It is easy to see that the test makes exactly k queries into v and accepts codewords of C with probability 1. The hard part of the analysis is to show that the test rejects non-codewords of C with noticeable probability. We do so in Section 5, for sufficiently large *odd* integer k provided C has small-bias. The resulting testing theorem (Theorem 5.5) is stated and proved in Section 5). When the code C is complementation closed however, $C^\perp$ has no codewords of odd weight. We extend our analysis to this setting by showing (in Section 6.1) that for sufficiently large *even* integer k, the canonical k-local test T_k is a strong k-local test for C. In Section 6.2, we consider codes of moderate bias (but large distance), and show that the canonical tests give local tests in this setting. This allows us to get (non-canonical) local tests for all sparse linear codes of large distance in Section 6.3.

3 Duals of sparse codes contain low weight codewords

Our approach (following [11]) is to use the "MacWilliams Transform" to estimate the size of $[C^\perp]_k$.

For a code C and integer $i \in \{0, \ldots, n\}$ let $B_i^C = |[C]_i|$ denote the number of codewords of weight i. We refer to the vector $\langle B_0^C, \ldots, B_n^C \rangle$ as the weight distribution of C. The MacWilliams Identities

show that the weight distribution of $C^\perp$ can be computed from the weight distribution of C. Specifically the relationship is as given below:

Theorem 3.1 (MacWilliams Identity, see e.g., [7]) *For a linear binary code C of length n, $B_j^{C^\perp} = \frac{1}{|C|}\sum_{i=0}^{n} B_i^C P_j(i)$, where $P_j(i) = P_j^n(i) = \sum_{\ell=0}^{j}(-1)^\ell \binom{i}{\ell}\binom{n-i}{j-\ell}$ is the Krawtchouk polynomial of degree j.*

In our case, we do not know the weight distribution of C, but we know a few things about it, which suffice to estimate the weight distribution of $C^\perp$. The following proposition summarizes our knowledge of the weight distribution of C and follows easily from the sparsity and the *bias* of the code.

Proposition 3.2 *The following hold for every n^t-sparse C of with distance $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$.*

1. $B_0^C = 1$.
2. $B_i^C = 0$ *for* $i \in \{1, \ldots, \frac{n}{2} - n^{1-\gamma}\}$.
3. $\sum_{i=0}^{n} B_i^C \leq n^t$.
4. *If C is $n^{-\gamma}$-biased then* $B_i^C = 0$ *for* $i \in \{\frac{n}{2} + n^{1-\gamma}, \ldots, n\}$.

The following proposition summarizes our knowledge of the Krawtchouk Polynomials, which is needed to translate the knowledge of the weight distribution of C into information about the weight distribution of $C^\perp$.

Proposition 3.3 *The following properties hold for Krawtchouk polynomials:*

1. $P_k(0) = \binom{n}{k}$.
2. *For every x, we have* $P_k(x) = (-1)^k P_k(n-x)$.
3. *[12, Equation (70)] $P_k(\cdot)$ has k real roots that lie between $n/2 - \sqrt{kn}$ and $n/2 + \sqrt{kn}$.*
4. *For $i \in [n]$, the following bounds hold*

$$\begin{array}{rcccl} 0 \leq & P_k(i) & \leq & (n-2i)^k/k! & \text{if} \quad 0 \leq i \leq n/2 - \sqrt{kn} \\ & |P_k(i)| & \leq & k^{k/2}n^{k/2}/k! & \text{if} \quad n/2 - \sqrt{kn} \leq i \leq n/2 + \sqrt{kn} \\ & P_k(i) & \leq & 0 & \text{if} \quad n/2 + \sqrt{kn} \leq i \leq n \text{ and } k \text{ is odd} \end{array}$$

5. $P_k(i-1) - P_k(i) = P_{k-1}(i) + P_{k-1}(i-1)$.
6. *For $i > n/2 + \sqrt{kn}$, $P_k(i)$ is decreasing in i for odd k, and increasing in i for even k.*

Proof: Parts (1) and (2) follow from the definition of $P_k(i)$. Part (3) is from [12, Equation (70)]. Part (4) follows from Part (3) and the definition of $P_k(i)$ by applying elementary facts about polynomials. Part (5) is from [12, Equation (17)]. Part (6) follows from Part (5) and the fact that $P_{k-1}(i)$ is positive in the given range for odd k and negative for even k. ■

We show next that the above already suffice to get a close estimate on $B_k^{C^\perp}$, the number of codewords in $C^\perp$ of weight k. Before doing so, we encapsulate a simple fact about sparse codes that will be used repeatedly in this paper: Roughly it states that the middle terms in the expression "$\frac{1}{|C|}\sum_{i=0}^{n} B_i^C P_k(i)$", for $B_k^{C^\perp}$, contribute a negligible amount compared to the first term. Though our principal concern at the moment are small-biased, sparse, codes; we prove the claim more generally, for any sparse set $S \subseteq \{0,1\}^n$.

Claim 3.4 *For every $\gamma > 0$ and $c, t < \infty$ if $k \geq (t+c+1)/\gamma$, then for any n^t-sparse set $S \subseteq \{0,1\}^n$, the following holds: $|\sum_{i=\frac{n}{2}-n^{1-\gamma}}^{\frac{n}{2}+n^{1-\gamma}} B_i^S P_k(i)| = o(n^{-c}) \cdot P_k(0)$. Furthermore, if k is odd, we have $\sum_{i=\frac{n}{2}-n^{1-\gamma}}^{n} B_i^S P_k(i) = o(n^{-c}) \cdot P_k(0)$.*

Proof: The proof is straightforward. Let $\alpha = 1 - \gamma$. We have:

$$\begin{aligned}
|\sum_{i=\frac{n}{2}-n^{\alpha}}^{\frac{n}{2}+n^{\alpha}} B_i^S P_k(i)| &\leq \sum_{i=\frac{n}{2}-n^{\alpha}}^{\frac{n}{2}+n^{\alpha}} B_i^S |P_k(i)| \\
&\leq \max_{\frac{n}{2}-n^{\alpha} \leq i_0 \leq \frac{n}{2}+n^{\alpha}} \{|P_k(i_0)|\} \sum_{i=\frac{n}{2}-n^{\alpha}}^{\frac{n}{2}+n^{\alpha}} B_i^S \\
&\leq 2n^{k\cdot\alpha} \sum_{i=\frac{n}{2}-n^{\alpha}}^{\frac{n}{2}+n^{\alpha}} B_i^S \text{ (Using Proposition 3.3, Part (4))} \\
&\leq 2n^{k\cdot\alpha} \sum_{i=0}^{n} B_i^S \\
&\leq 2n^{k\cdot\alpha}|S| \\
&= o(n^{-c})n^k = o(n^{-c})P_k(0)
\end{aligned}$$

where the last line uses the fact that $k \geq (t+c+1)/\gamma$. The first part of the claim follows immediately. The second part then follows from the fact that $P_k(i) < 0$ for every $i \in \{\frac{n}{2} + n^{1-\gamma}, \ldots, n\}$. ■

Lemma 3.5 *Let C be an n^t-sparse, code with $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$. Then for every c, t, γ, there exists a k_0 such that for every odd $k \geq k_0$ $B_k^{C^\perp} \leq \frac{P_k(0)}{|C|} \cdot (1 + o(n^{-c}))$. If C is $n^{-\gamma}$-biased then for every (odd and even) $k \geq k_0$ it is the case that $B_k^{C^\perp} = \frac{P_k(0)}{|C|} \cdot (1 + \theta(n^{-c}))$.*

Remark: Here and later we use the notation $f(n) = g(n) + \theta(h(n))$ to imply that for every ϵ and for sufficiently large n $g(n) - \epsilon \cdot h(n) \leq f(n) \leq g(n) + \epsilon \cdot h(n)$.

Proof: By the MacWilliams Identities (Theorem 3.1) we have:

$$\begin{aligned}
B_k^{C^\perp} &= \frac{1}{|C|}\sum_{i=0}^{n} B_i^C P_k(i) \\
&= \frac{P_k(0)}{|C|} + \frac{1}{|C|}\sum_{i=1}^{n} B_i^C P_k(i) \\
&= \frac{P_k(0)}{|C|} + \frac{1}{|C|}\sum_{i=\frac{n}{2}-n^{1-\gamma}}^{n} B_i^C P_k(i) \quad \text{(Using Proposition 3.2, Part (2))}
\end{aligned}$$

We can now use Claim 3.4 to see that the latter summation is upper bounded by $o(n^{-c}).P_k(0)/|C|$. The first part of the lemma then follows. For the second part, where C is $n^{-\gamma}$-biased, we see that

$$\begin{aligned} B_k^{C^\perp} &= \frac{P_k(0)}{|C|} + \frac{1}{|C|}\sum_{i=\frac{n}{2}-n^{1-\gamma}}^{n} B_i^C P_k(i) \\ &= \frac{P_k(0)}{|C|} + \frac{1}{|C|}\sum_{i=\frac{n}{2}-n^{1-\gamma}}^{\frac{n}{2}+n^{1-\gamma}} B_i^C P_k(i) \quad \text{(Using Proposition 3.2, Part (4))} \\ &= \frac{P_k(0)}{|C|} + \theta(n^{-c})\frac{P_k(0)}{|C|} \quad \text{(Using Claim 3.4),} \end{aligned}$$

which yields the second part of the lemma statement. ■

4 Analysis of the self-corrector

We start by proving that the self-corrector of Section 2 works correctly. Specifically we prove Theorem 1.2 by showing that for sufficiently large k, the Self-Corrector SC_k of Section 2 correctly computes x_i on input i and oracle access to $v \in \{0,1\}^n$ that is δ-close to $x \in C$, provided $\delta < \frac{1}{2k}$.

To analyze the self-corrector, we study the distribution of the bits y_i, y_j when y is a randomly chosen codeword of $C^\perp$ of weight k. Now, if the codewords of $[C^\perp]_k$ had been uniformly chosen from the vectors of weight k, then y_i and y_j would be almost 2-wise independent (but not exactly so, since we are sampling two items without replacement from a universe of size n). We show below that the distribution of y_i, y_j is almost the same when y is chosen uniformly from $[C^\perp]_k$, for any sparse code C of low bias. (see Lemma 4.3).

To understand the distribution of y_i, we need to approximate the size of the set $[C^\perp]_{k,i}$. (Recall that $[C^\perp]_{k,i} = \{y \in [C^\perp]_k | y_i = 1\}$.) We do so by looking at the code C^{-i} which is the code C with the ith coordinate punctured. In other words, $C^{-i} = \{\pi_{-i}(x) | x \in C\}$, where $\pi_{-i} : \{0,1\}^n \to \{0,1\}^{n-1}$ is the projection function $\pi_{-i}(y_1, \ldots, y_n) = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. The following proposition explains the relevance of this projection to $[C^\perp]_{k,i}$.

Proposition 4.1 *Let* $\pi_{-i}^{-1}(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n) = (y_1, \ldots, y_{i-1}, 0, y_{i+1}, \ldots, y_n)$ *and* $\pi_{-i}^{-1}(S) = \{\pi_{-i}^{-1}(y) | y \in S\}$. *Then* $[C^\perp]_{k,i} = [C^\perp]_k - \pi_{-i}^{-1}([(C^{-i})^\perp]_k)$, *and thus* $|[C^\perp]_{k,i}| = |[C^\perp]_k| - |[(C^{-i})^\perp]_k|$.

Proof: Note that if $y \in C^\perp$ with $y_i = 0$ then $\pi_{-i}(y) \in (C^{-i})^\perp$. So we have $\{\pi_{-i}(y) | y \in C^\perp, y_i = 0\} \subseteq (C^{-i})^\perp$ and we claim that actually $(C^{-i})^\perp = \{\pi_{-i}(y) | y \in C^\perp, y_i = 0\}$. This is easy to see by counting. On the one hand, $|(C^{-i})^\perp| = 2^{n-1}/|C^{-i}| = 2^{n-1}/|C| = \frac{1}{2}|C^\perp|$ (where the equality $|C^{-i}| = |C|$ uses the fact that the (non-normalized) Hamming distance of C is at least 2). On the other hand, $|\{\pi_{-i}(y) | y \in C^\perp, y_i = 0\}| \geq \frac{1}{2}|C^\perp|$ since at least half the vectors in $C^\perp$ must have the ith coordinate set to 0.

It immediately follows that $[C^\perp]_{k,i} = [C^\perp]_k - \pi_{-i}^{-1}([(C^{-i})^\perp]_k)$. ■

Extending our notation to consider codes punctured twice, let $\pi_{-\{i,j\}}$ denote the projection to all coordinates except i and j, and let $C^{-\{i,j\}}$ denote the projection of the code C by $\pi_{-\{i,j\}}$. Let $[C^\perp]_{k,\{i,j\}} = \{y \in [C^\perp]_k | y_i = y_j = 1\}$. Then, we get:

Proposition 4.2 *For every $i \neq j$, $[C^{\perp}]_{k,\{i,j\}} = [C^{\perp}]_k - \pi_{-i}^{-1}([(C^{-i})^{\perp}]_k) - \pi_{-j}^{-1}([(C^{-j})^{\perp}]_k) + \pi_{-\{i,j\}}^{-1}([(C^{-\{i,j\}})^{\perp}]_k)$. Thus $|[C^{\perp}]_{k,\{i,j\}}| = |[C^{\perp}]_k| - |[(C^{-i})^{\perp}]_k| - |[(C^{-j})^{\perp}]_k| + |[(C^{-\{i,j\}})^{\perp}]_k|$.*

Proof: Follows from Proposition 4.1 applied to C^{-j}, and by inclusion-exclusion. ■

Armed with the facts above, and our knowledge of the weight distribution of duals of sparse small-biased codes, it is easy to prove the following lemma.

Lemma 4.3 *For every $\gamma > 0$ and $c, t < \infty$ there exists k such that for sufficiently large n the following holds: Let $C \subseteq \{0,1\}^n$ be an n^t-sparse $n^{-\gamma}$-biased linear code. Then for every $i \neq j \in [n]$, the probability that $y_j = 1$ for a randomly chosen vector $y \in [C^{\perp}]_{k,i}$ is $(k-1)/(n-1) + \theta(n^{-c})$.*

Proof: Note that the quantity we are interested in is the ratio $|[C^{\perp}]_{k,\{i,j\}}|/|[C^{\perp}]_{k,i}|$. By Propositions 4.1 and 4.2 it suffices to get good estimates on $|[C^{\perp}]_k|$, $|[(C^{-i})^{\perp}]_k|$, $|[(C^{-j})^{\perp}]_k|$, and $|[(C^{-\{i,j\}})^{\perp}]_k|$. Fortunately, this is easy, since C, C^{-i}, C^{-j} and $C^{-\{i,j\}}$ are all n^t-sparse, $n^{-\gamma}$-biased, codes. The only difference is their lengths, which are n, $n-1$, $n-1$ and $n-2$ respectively. Applying Lemma 3.5 to these in turn, we find that there exists a k such that

$$\begin{aligned} |[C^{\perp}]_k| &= \binom{n}{k} + \theta(n^{k-(c+2)}) \\ |[(C^{-i})^{\perp}]_k| &= \binom{n-1}{k} + \theta(n^{k-(c+2)}) \\ |[(C^{-j})^{\perp}]_k| &= \binom{n-1}{k} + \theta(n^{k-(c+2)}) \\ \text{and } |[(C^{-\{i,j\}})^{\perp}]_k| &= \binom{n-2}{k} + \theta(n^{k-(c+2)}). \end{aligned}$$

By Proposition 4.1, we have $|[C^{\perp}]_{k,i}| = \binom{n-1}{k-1} + \theta(n^{k-(c+2)})$ and, by Proposition 4.2, we have $|[C^{\perp}]_{k,\{i,j\}}| = \binom{n-2}{k-1} + \theta(n^{k-(c+2)})$. We conclude that $|[C^{\perp}]_{k,\{i,j\}}|/|[C^{\perp}]_{k,i}| = (k-1)/(n-1) + \theta(n^{-c})$. ■

We now state and prove our main lemma, which immediately implies the correctness of the self-corrector SC_k.

Lemma 4.4 *For every $t < \infty$, $\gamma > 0$ there exists a $k = k_{t,\gamma} < \infty$ such that the following holds: Let C be an n^t-sparse, $n^{-\gamma}$-biased code. Let $v \in \{0,1\}^n$ be δ-close to $x \in C$. Then for every $i \in [n]$, $\Pr[\mathrm{SC}_k^v(i) \neq x_i] \leq k\delta + \theta(1/n)$.*

Proof: Pick k large enough to be able to apply Lemma 4.3 for $c = 2$. Thus, for random $y \in [C^{\perp}]_{k,i}$ and $j \neq i$, we have $\Pr_y[y_j = 1] \leq (k-1)/(n-1) + \theta(n^{-2}) \leq k/n + \theta(n^{-2})$.

Let $E = \{j \in [n] | v_j \neq x_j\}$. We have $|E| \leq \delta n$. Now consider $y \in_U [C^{\perp}]_{k,i}$ and let $S_y = \{j \in [n] - \{i\} | y_j = 1\}$. If $S_y \cap E = \emptyset$, we have $\mathrm{SC}_k^v(i) = x_i$. Thus it suffices to bound the probability $S_y \cap E \neq \emptyset$ from above. We bound this by $|E| \cdot \max_{j \in E}\{\Pr_y[y_j = 1]\} \leq k\delta + \theta(n^{-1})$. ■

Theorem 1.2 now follows, for k as given by Lemma 4.4, where the parameter τ in the definition of the self-corrector (the fraction or errors that can be tolerated can be chosen to be less than $\frac{1}{2k}$ and the error probability ϵ being $k\tau + o(1/n)$.

4.1 Local Decodability of Sparse Codes

In this section we consider a "low-biased" linear encoding function $E : \{0,1\}^{t \log n} \to \{0,1\}^n$ and show that it is locally decodable with high probability. By linear encoding we mean that $E(m) = A \cdot m$ for some matrix $A \in \{0,1\}^{n \times (t \log n)}$. Low-biased implies that the code $C = \{E(m) | m \in \{0,1\}^{t \log n}\}$ is $n^{-\gamma}$-biased for some $\gamma > 0$. Local decodability of E follows easily from the following proposition.

Proposition 4.5 *Let E be a linear map and $C = \{E(m) | m \in \{0,1\}^{t \log n}\}$ be an $n^{-\gamma}$-biased code. Let C' be the linear code $C' = \{(m, E(m)) | m \in \{0,1\}^{t \log n}\}$. Then C' is n^t-sparse, $O(n^{-\gamma})$-biased, and given oracle access to a word $v \in \{0,1\}^n$ such that $\delta(v, E(m)) \leq \delta$, one can get oracle access to a word $\tilde{v} \in \{0,1\}^{t \log n + n}$ that is $(\delta + \frac{t}{n} \log n)$-close to the codeword $(m, E(m)) \in C'$.*

Proof: All parts are immediate. For the last part, we use an oracle for $\tilde{v} = (0^{t \log n}, v)$. ■ By Theorem 1.2, the code C' above is self-correctible and so E is locally-decodable, and so we get:

Corollary 4.6 *For every $t < \infty$ and $\gamma > 0$ there exists a k such that the following holds. Let E be an $n^{-\gamma}$-biased linear encoding mapping $\{0,1\}^{t \log n}$ to $\{0,1\}^n$. Then E is k-locally decodable.*

5 Analysis of the local test

The results of the previous sections show that $C^\perp$ does have many low-weight codewords and that they suffice to give good local error-correction algorithms for C. Yet they do not even imply that the low-weight codewords of $C^\perp$ specify C uniquely. For all we know the set of vectors $\{x \in \{0,1\}^n | \langle x, y \rangle = 0 \ \forall y \in [C^\perp]_k\}$ may be a superset of C. In this section we show that this does not happen. The low weight codewords of $C^\perp$ completely define C in the sense that $C = \{x | \langle x, y \rangle = 0 \ \forall y \in [C^\perp]_k\}$ for some constant k. Furthermore, we show that the canonical tester of Section 2 rejects words at large distance from the code with high probability.

To derive these results, we fix a word $v \in \{0,1\}^n$ that does not belong to C and attempt to understand two structures. First we consider the linear code $C||v = C \cup (C + v)$ where $C + v = \{x + v | x \in C\}$. We then turn to the new set of words $C + v$ which form a coset of the code C and attempt to understand it. Together these sets help us understand the performance of the test of Section 2.

To see how these sets become relevant, we start with a simple proposition relating the probability that the test test T_k^v rejects to the code $C||v$.

Proposition 5.1

$$\mathrm{Rej}_k(v) = 1 - \frac{B_k^{(C||v)^\perp}}{B_k^{C^\perp}}.$$

Proof: The proposition follows from the fact that the test T_k^v accepts on random choice $y \in [C^\perp]_k$ if and only if $y \in [(C||v)^\perp]_k$ (since the latter containment happens if and only if $\langle v, y \rangle = 0$). ■

In what follows we show that $\mathrm{Rej}_k(v) > 0$ if $v \notin C$. In fact, we show the stronger result that $\mathrm{Rej}_k(v) = \Omega(\delta(v, C))$.

Lemma 5.2 *For every $c, t < \infty$ and $\gamma > 0$ there exists a k_0 such that the following hold: Let C be an n^t-sparse code of distance $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$. Let $v \in \{0,1\}^n \setminus C$ be δ-far from C (i.e., $\delta(v, C) = \delta$). Then, for odd $k \geq k_0$,*

$$B_k^{(C||v)^\perp} \leq \left(1 - \delta/2 + o(n^{-c})\right) \cdot \frac{P_k(0)}{|C|}.$$

Proof: We prove the lemma using two "sublemmas" (Lemmas 5.3 and 5.4) stated and proved below, which bound some technical terms that will appear in this proof.

Let $\gamma' = \gamma/2$. We prove the lemma for $k_0 = \max\{k_1, k_2, 400\}$ where k_1 is chosen to be big enough so that Claim 3.4 applies, and k_2 is the constant given by Lemma 3.5, for t, c and γ'.

Using the MacWilliams Identities (Theorem 3.1) we have

$$\begin{aligned}
B_k^{(C||v)^\perp} &= \frac{1}{|C||v|} \sum_{i=0}^{n} B_i^{(C||v)} P_k(i) \\
&= \frac{1}{2|C|} \sum_{i=0}^{n} B_i^{C} P_k(i) + \frac{1}{2|C|} \sum_{i=0}^{n} B_i^{C+v} P_k(i) \quad (\text{Since } (C||v) = C \cup C + v) \\
&= \frac{1}{2} B_k^{C^\perp} + \frac{1}{2|C|} \sum_{i=0}^{n} B_i^{C+v} P_k(i).
\end{aligned}$$

By Lemma 3.5 we have that $B_k^{C^\perp} \leq \frac{P_k(0)}{|C|} \cdot (1 + o(n^{-c}))$ and so to prove the lemma it suffices to show that

$$\frac{1}{|C|} \sum_{i=0}^{n} B_i^{C+v} P_k(i) \leq (1 - \delta + o(n^{-c})) \cdot \frac{P_k(0)}{|C|}.$$

Since $C+v$ is an n^t-sparse code of distance $\frac{1}{2} - n^{-\gamma}$, by Claim 3.4 we have $\sum_{i=\frac{n}{2}-n^{1-\gamma'}}^{n} B_i^{C+v} P_k(i) = o(n^{-c} \cdot P_k(0))$. It suffices to prove that

$$\sum_{i=0}^{\frac{n}{2}-n^{1-\gamma'}} B_i^{C+v} P_k(i) \leq (1 - \delta + o(n^{-c})) \cdot P_k(0).$$

We now know that $B_i^{C+v} = 0$ for every $i = \{0, \frac{n}{2} - n^{1-\gamma'} - \delta n\}$ except possibly for $i = \delta n$. If $\delta n < \frac{n}{2} - n^{1-\gamma'} - \delta n$ we are within the "unique" decoding radius and so $B_{\delta n}^{C+v} = 1$. Thus we have

$$\sum_{i=0}^{\frac{n}{2}-n^{-\gamma'}} B_i^{C+v} P_k(i) \leq P_k(\delta n) + \sum_{i=a}^{b} B_i^{C+v} P_k(i),$$

where $a = \max\{\frac{n}{2} - n^{1-\gamma'} - \delta n, \delta n\}$ and $b = \frac{n}{2} - n^{1-\gamma'}$ In Lemma 5.3 below we show the first term is at most $(1-\delta)^k \cdot P_k(0)$. In Lemma 5.4 below we show the second term is $12 P_k(0) \cdot \left(\min\{(1-2\delta)^{k-2}, (4\delta)^{k-2}\} + n^{-(k-2)\gamma'}\right)$. We thus conclude that

$$\sum_{i=0}^{\frac{n}{2}-n^{-\gamma'}} B_i^{C+v} P_k(i) \leq P_k(0) \cdot \left((1-\delta)^k + 12 \min\{(1-2\delta)^{k-2}, (4\delta)^{k-2}\} + o(n^{-c})\right).$$

We claim that the expression $(1-\delta)^k + 12\min\{(1-2\delta)^{k-2}, (4\delta)^{k-2}\} \leq 1-\delta$ for every $0 \leq \delta \leq \frac{1}{2}$. For $\delta \leq \frac{1}{100}$ and $k \geq 6$, we have that $(1-\delta)^k + 12\cdot(4\delta)^{k-2} \leq 1-(k/2)\delta + 12(4\delta)^2 \leq 1-3\delta+2\delta \leq 1-\delta$. For $\frac{1}{100} \leq \delta \leq \frac{1}{2}$ and $k \geq 400$, we have $(1-\delta)^k \leq \frac{1}{4}$ and $12(1-2\delta)^{k-2} \leq \frac{1}{4}$. So $(1-\delta)^k + 12(1-2\delta)^{k-2} \leq \frac{1}{2} \leq 1-\delta$. The lemma now follows. ■

We now present the "sublemmas".

Lemma 5.3 *For every k, for sufficiently large n (i.e., $n \geq n_k$) and for every $\delta < 1/2$, $P_k(\delta n) \leq (1-\delta)^k \cdot P_k(0)$.*

Proof: We consider two cases on the value of δ.

If $\delta \leq \frac{1}{2k}$, we claim that $P_k(\delta n) \leq \binom{n-\delta n}{k}$. To see this let $i = \delta n$ and $T_\ell = \binom{i}{\ell}\binom{n-i}{k-\ell}$, so that $P_k(i) = \sum_{\ell=0}^{k}(-1)^\ell T_\ell$. We note that under the restriction on δ, we have $T_\ell \geq T_{\ell+1}$ for every ℓ and so the alternating sum $\sum_{\ell=0}^{k}(-1)^\ell T_\ell$ can be upper bounded by its first term $T_0 = \binom{n-i}{k}$. It is now elementary to see that $T_0 \leq (1-\delta)^k P_k(0)$.

For the case that $\delta \geq \frac{1}{2k}$ we use the fact (from Proposition 3.3, Part (4)) that $P_k(\delta n) \leq (1-2\delta)^k n^k/k!$. Writing $n^k/k! = \binom{n}{k} + O(n^{k-1})$, we get $P_k(\delta n) \leq (1-2\delta)^k(P_k(0) + O(n^{k-1}))$. For sufficiently large n it is easy to see that this is at most $(1-\delta)^k \cdot P_k(0)$ for every $\delta \geq \frac{1}{2k}$. ■

Lemma 5.4 *Let k, t, γ be constants. Let $\gamma' \leq \gamma/2$. For sufficiently large n, let D be an n^t-sparse code of distance at least $\frac{1}{2} - n^{-\gamma}$. Let $\delta \leq \frac{1}{2}$, and let $a = \max\{\frac{n}{2} - n^{1-\gamma'} - \delta n, \delta n\}$ and $b = \frac{n}{2} - n^{1-\gamma'}$. Then*

$$\sum_{i=a}^{b} P_k(i) B_i^D = 12 P_k(0) \cdot \left(\min\{(1-2\delta)^{k-2}, (4\delta)^{k-2}\} + n^{-(k-2)\gamma'}\right).$$

Proof: The crux of this proof is the Johnson bound which says that the number of codewords in a ball of radius $i \leq n/2 - n^{1-\gamma/2}$ is at most $2n^2/(n-2i)^2$, for a code of (relative) distance $\frac{1}{2} - n^{-\gamma}$.

Let $m_i = 2n^2/(n-2i)^2$. Then, by the Johnson bound we have that $\sum_{j=0}^{i} B_i^D \leq m_i$ for every $i \leq b$. We first note that

$$\sum_{i=a}^{b} P_k(i) B_i^D \leq \frac{1}{k!}\sum_{i=a}^{b}(n-2i)^k B_i^D \leq \frac{1}{k!}(n-2a)^k m_a + \frac{1}{k!}\sum_{i=a+1}^{b}(n-2i)^k \cdot (m_i - m_{i-1}).$$

(The last inequality follows from the elementary fact that for non-negative numbers $x_1, \ldots, x_\ell, y_1, \ldots, y_\ell$ and $z_1 \geq z_2 \geq \cdots \geq z_\ell \geq 0$, if $\sum_{j=1}^{i} x_i \leq \sum_{j=1}^{i} y_i$ for every $i \in [\ell]$, then $\sum_{i=1}^{\ell} x_i z_i \leq \sum_{i=1}^{\ell} y_i z_i$. We apply this fact to $x_i = B_i^D$ and $y_i = m_i - m_{i-1}$ and $z_i = (n-2i)^k$.)

We thus turn to upper bounding $\frac{1}{k!}(n-2a)^k m_a + \frac{1}{k!}\sum_{i=a+1}^{b}(n-2i)^k \cdot (m_i - m_{i-1})$. Looking at the first term, we have

$$(n-2a)^k/k! m_a \leq (n-2a)^k/k! \cdot (2n^2)/(n-2a)^2 = 2n^2(n-2a)^{k-2}/k!$$

To analyze the second term, note that $m_i - m_{i-1} \leq 8n^2/(n-2i)^3$. A straightforward calculation now yields $\frac{1}{k!}\sum_{i=a+1}^{b}(n-2i)^k \cdot (m_i - m_{i-1}) \leq \frac{4n^2}{k!}(n-2a)^{k-2}$. Thus, we have

$$\frac{1}{k!}\sum_{i=a+1}^{b}(n-2i)^k \cdot (m_i - m_{i-1}) \quad \leq \quad \frac{1}{k!}\sum_{i=a+1}^{b}(n-2i)^k \cdot 8n^2/(n-2i)^3$$

$$\begin{aligned} &\leq \frac{8n^2}{k!}\sum_{i=a+1}^{b}(n-2i)^{k-3} \\ &\leq \frac{8n^2}{k!}(b-a)(n-2a)^{k-3} \\ &\leq \frac{4n^2}{k!}(n-2a)^{k-2}. \end{aligned}$$

Thus we conclude that $\sum_{i=a}^{b} P_k(i)B_i^D \leq 6n^2(n-2a)^{k-2}/k!$. Substituting the given expression for a, and using the crude bound $n^k/k! \leq 2P_k(0)$ we get

$$\sum_{i=a}^{b} P_k(i)B_i^D \leq 12P_k(0)\cdot\min\{(1-2\delta)^{k-2}, (4\delta)^{k-2} + (2n^{1-\gamma'})^{k-2}\}.$$

■

We are now ready to state and prove our main testing theorem for small-biased codes.

Theorem 5.5 *For every $t < \infty$ and $\gamma > 0$ there exists a constant $k = k_{t,\gamma} < \infty$ such that the following holds: If $C \subseteq \{0,1\}^n$ is n^t-sparse and $n^{-\gamma}$-biased, then C is strongly k-locally testable.*

Proof: Given t and γ, let k be an odd integer greater than k_0 as given by Lemma 5.2 for $c = 2$. We claim that the tester T_k has the required properties. As observed in Section 2 the test makes k queries and accepts codewords of C. By Proposition 5.1, for $v \notin C$ it rejects with probability (exactly) $1 - B_k^{(C||v)^\perp}/B_k^{C^\perp}$. Let $\delta = \delta(v, C)$. By Lemma 5.2, we have $B_k^{(C||v)^\perp} \leq \frac{P_k(0)}{|C|}\cdot(1-\delta/4+o(n^{-2}))$, while by Lemma 3.5 we have $B_k^{C^\perp} = \frac{P_k(0)}{|C|}\cdot(1+\theta(n^{-2}))$. Thus the rejection probability is at least $\delta/4 - o(n^{-2})$. Using $\delta \geq \frac{1}{n}$, we thus get that for sufficiently large n, the rejection probability is at least $\delta/8$. We thus have that T_k satisfies the definition of a k-local strong tester with $\epsilon' = \frac{1}{8}$. ■

6 Extensions to non-small-biased codes

In this section we first extend the results (easily) to the class of complementation-closed codes. We then consider a class of "moderately-biased" high distance codes, and give a weak tester for such codes. Finally, we give a generic reduction from testing high-distance codes (of arbitrary bias) to high-distance codes of moderate bias. This gives us a weak tester for all sparse codes of large distance.

6.1 Testing and Correcting Complementation-Closed Codes

Recall that here we consider codes of the form $C = C'||1^n$ and show how to test and correct them. The analysis is a straightforward modification of the analyses of Sections 3, 4 and 5. The main difference now is that $C^\perp$ has no odd weight codewords! But we also know that $B_i^C = B_{n-i}^C$ and also $B_i^{(C||v)} = B_{n-i}^{(C||v)}$ for every $v \in \{0,1\}^n$. This allows use to study only the "bottom" half of

sums in the MacWilliams Identities and then conclude that the top half behaves similarly. (E.g., $\sum_{i=0}^{n} B_i^C P_k(i) = 2\sum_{i=0}^{n/2} B_i^C P_k(i)$ etc.)
Specifically, we get the following variation of Lemma 3.5.

Lemma 6.1 *For every c, t, γ, there exists a k_0 such that for every* even *$k \geq k_0$, $B_k^{C^\perp} = 2 \cdot P_k(0) \cdot (1 + \theta(n^{-c}))$.*

Using this lemma we can now prove Lemma 4.4 for the class of complementation-closed codes as well, and thus get a self-corrector for such codes. To get a local tester, we prove the following variation of Lemma 5.2.

Lemma 6.2 *For every $c, t < \infty$ and $\gamma > 0$ there exists a k_0 such that the following hold: Let C be an n^t-sparse $n^{-\gamma}$-biased code. Let $v \in \{0,1\}^n \setminus C$ be δ-far from C (i.e., $\delta(v, C) = \delta$). Then, for* even *$k \geq k_0$,*

$$B_k^{(C||v)^\perp} \leq \left(1 - \delta/2 + o(n^{-c})\right) \cdot B_k^{C^\perp}.$$

The analysis of the local test follows, and thus we get Theorem 1.3.

6.2 Testing moderate-bias codes

Before moving on to general sparse, high-distance codes, we consider the case where a code does not have codewords of weight greater than say, $\frac{5}{6}$ (while the distance of the code is still very close to half, specifically, $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$). We start with a simple, but weak analysis, for a local test for such codes.

Lemma 6.3 *For every $\gamma > 0, t$ there exist functions $q : (0,1] \to \mathbb{Z}^+$ and $\epsilon : (0,1] \to (0,1]$, with $q(\delta) = O(\log \frac{1}{\delta})$ and $\epsilon(\delta) = \Omega(\delta)$, such that the following holds: If C is a 1/3-biased, n^t-sparse code of distance $\frac{1}{2} - n^{-\gamma}$, then C is (weakly) (q, ϵ)-locally testable.*

Proof: Fix the parameter $\delta > 0$ and let $w \in \{0,1\}^n$ satisfy $\delta(w, C) \geq \delta$. As usual we look at the quantities $B_k^{C^\perp}$ and $B_k^{(C||w)^\perp}$ for sufficiently large odd integer k. By Lemma 5.2 we have that for sufficiently large odd $k \geq k_{\gamma,t}$,

$$B_k^{(C||w)^\perp} \leq (1 - \delta/2 - \theta(n^{-c})) \cdot P_k(0)/|C|.$$

We now turn to bounding $B_k^{C^\perp}$ from below. (The following is just a minor modification of the proof of Lemma 5.2.) As usual we have

$$\begin{aligned}
B_k^{C^\perp} &= \frac{1}{|C|} \sum_{i=0}^{n} B_i^C P_k(i) \\
&= \frac{P_k(0)}{|C|} + \frac{1}{|C|} \sum_{i=n/2-n^{1-\gamma}}^{n/2+n^{1-\gamma}} B_i^C P_k(i) + \frac{1}{|C|} \sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i) \\
&= \frac{P_k(0)}{|C|} \cdot (1 + \theta(n^{-c})) + \frac{1}{|C|} \sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i) \quad \text{(Using Claim 3.4)}
\end{aligned}$$

We thus turn to lower bounding the final term $\sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i)$ above. Note that this term may be negative and we wish to prove it is not "too" negative. Below we'll upper bound the absolute value of this summation. Let τ be such that the weight of the maximum weight codeword in C is $(1-\tau)n$. Note that $\tau \geq 1/6$. Then $|\sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i)| \leq |P_k((1-\tau)n)| + |\sum_{i=n/2+n^{1-\gamma}}^{n-a} B_i^C P_k(i)|$ where $a = \max\{(\tau n, n/2 - n^{1-\gamma} - \tau n)\}$. Using the fact that $|P_k(i)| = |P_k(n-i)|$, we find thus that $|\sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i)| \leq P_k(\tau n) + \sum_{i=a}^{b} B_i^C P_k(i)$ where $b = \frac{n}{2} - n^{1-\gamma}$. We are now in a position to apply Lemmas 5.3 and 5.4. By Lemma 5.3, we see that the first term is at most $(1-\tau)^k \cdot P_k(0)$, while by Lemma 5.4, the second term is at most $12P_k(0) \cdot \left((1-2\tau)^{k-2} + n^{-(k-2)\gamma}\right)$. Using $\tau \geq \frac{1}{6}$ and setting $k = \Omega(\log \frac{1}{\delta})$, we get that $|\sum_{i=n/2+n^{1-\gamma}}^{\frac{5}{6}n} B_i^C P_k(i)| \leq P_k(0) \cdot \left(\frac{\delta}{10} + \theta(n^{-c})\right)$. We conclude that for such a choice of k, $B_k^{C^\perp} \geq \frac{P_k(0)}{|C|} \cdot \left(1 - \frac{\delta}{10} - \theta(n^{-c})\right)$.

Putting the bounds on B_k^C and $B_k^{(C||w)}$ together, and applying Proposition 5.1, we find that the canonical test T_k, is a $(O(\log \frac{1}{\delta}), \Omega(\delta))$-local test for C. ■

We now give a more complicated analysis, but now yielding a strong local test for moderate codes.

Lemma 6.4 *For every $t < \infty, \gamma > 0$ there exists a $k = k_{t,\gamma} < \infty$ such that the following holds: If C is a $1/3$-biased, n^t-sparse code of distance $\frac{1}{2} - n^{-\gamma}$, then C is strongly k-locally testable.*

Proof: As usual we use a canonical tester T_k for a sufficiently large (but constant, given t and γ) odd integer k. Below we argue that such a test rejects non-codewords with probability proportional to their distance from the code.

Let $w \in \{0,1\}^n$ be such that $\delta(w, C) = \delta$. For an appropriate choice of k, we wish to bound the quantity $1 - B_k^{(C||w)^\perp}/B_k^{C^\perp}$. While in previous analyses, we bounded the two quantities separately, this time we will work with the two quantities together.

Let $b = \frac{n}{2} + n^{1-\gamma}$. First note that as usual we have

$$B_k^{C^\perp} = \frac{P_k(0)}{|C|} \cdot (1 + \theta(n^{-c})) + \frac{1}{|C|} \cdot \left(\sum_{i=b}^{n} P_k(i) \cdot B_i^C\right).$$

Note that the final term above is negative for odd k, and so poses a problem in our analysis. In this proof, we overcome this obstacle by showing that a similar negative contribution occurs in the expression for $B_k^{(C||w)^\perp}$. We have

$$B_k^{(C||w)^\perp} \leq \frac{B_k^{C^\perp}}{2} + \frac{P_k(0)}{2|C|} \cdot (1 - \delta + \theta(n^{-c})) + \frac{1}{2|C|} \sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w}.$$

By Proposition 5.1 the rejection probability, which we'll denote by ρ, is lower bounded by

$$\begin{aligned} \rho &\geq 1 - B_k^{(C||w)^\perp}/B_k^{C^\perp} \\ &\geq 1 - \frac{1}{2} - \frac{P_k(0)}{2|C| \cdot B_k^{C^\perp}} \cdot (1 - \delta + \theta(n^{-c})) + \frac{1}{2|C| \cdot B_k^{C^\perp}} \sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w} \end{aligned}$$

Let $\alpha = \frac{\sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w}}{P_k(0)}$ and $\beta = \frac{\sum_{i=b}^{n} P_k(i) \cdot B_i^{C}}{P_k(0)}$. (Note both quantities are negative!) Then we have

$$\begin{aligned}
\rho &\geq \frac{1}{2} - \frac{P_k(0)}{2|C| \cdot B_k^{C^\perp}} \cdot (1 - \delta + \theta(n^{-c})) + \frac{1}{2|C| \cdot B_k^{C^\perp}} \sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w} \\
&= \frac{1}{2} - \frac{1 - \delta + \theta(n^{-c}) + \alpha}{2(1 + \theta(n^{-c}) + \beta)} \\
&= \frac{1 + \theta(n^{-c}) + \beta - 1 + \delta + \theta(n^{-c}) - \alpha}{2(1 + \theta(n^{-c}) + \beta)} \\
&\geq \frac{\delta + \beta - \alpha + \theta(n^{-c})}{2}
\end{aligned}$$

Using Lemma 6.5 below with $\epsilon = \frac{1}{2}$, we get $\beta - \alpha \geq -\frac{\delta}{2}$ and thus $\rho \geq \delta/4 + \theta(n^{-c})$. For sufficiently large n (using say $c = 1$) we get $\rho \geq \delta/8$, thus showing that for odd k, the test T_k is a strong k-local test for $\epsilon' = 1/8$. ■

The following lemma bounds the expression encountered in the proof of the previous lemma.

Lemma 6.5 *For every $t < \infty, \gamma, \epsilon > 0$, there exists a $k_0 = k_{t,\gamma,\epsilon}$ such that the following holds for every odd $k \geq k_0$: Let C be an n^t-sparse, $\frac{1}{3}$-biased code of distance $\delta(C) \geq \frac{1}{2} - n^{-\gamma}$. Let $w \in \{0,1\}^n$ and $\delta > 0$ be such that $\delta(w, C) = \delta n$. Then*

$$\sum_{i=b}^{n} P_k(i) \cdot B_i^{C} - \sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w} \geq -\epsilon \cdot \delta \cdot P_k(0),$$

where $b = \frac{n}{2} + n^{1-\gamma}$.

Proof: Without loss of generality, we assume that $\mathrm{wt}(w) = \delta \cdot n$. (If not, we can work with some $\tilde{w} \in C + w$ with weight δn, since $C||\tilde{w} = C||w$.) We also assume that $\gamma < \frac{1}{2}$. If not we prove the lemma for some $b' = \frac{n}{2} + k\sqrt{n}$ and $\epsilon' = \epsilon/2$. We can then easily lower bound the difference $\sum_{i=b}^{b'} P_k(i) \cdot B_i^{C} - \sum_{i=b}^{b'} P_k(i) \cdot B_i^{C+w}$ by $-\epsilon/2\delta P_k(0)$ using Claim 3.4, thus yielding the lemma (for $k = k_{t,\gamma',\epsilon'}$).

Let $S_1 = \{x \in C \mid b + \delta n \leq \mathrm{wt}(x) \leq \frac{5}{6}n\}$ and let $S_2 = \{x \in C \mid b \leq \mathrm{wt}(x) < b + \delta n\}$. Since $P_k(i)$ is negative for the range of interest, we have

$$\sum_{i=b}^{n} P_k(i) \cdot B_i^{C} - \sum_{i=b}^{n} P_k(i) \cdot B_i^{C+w} \geq \sum_{x \in S_1} (P_k(\mathrm{wt}(x)) - P_k(\mathrm{wt}(x+w))) + \sum_{x \in S_2} P_k(\mathrm{wt}(x).$$

Let $T_1 = \sum_{x \in S_1} (P_k(\mathrm{wt}(x)) - P_k(\mathrm{wt}(x+w)))$ denote the first term above, and let $T_2 = \sum_{x \in S_2} P_k(\mathrm{wt}(x))$ denote the latter quantity. We bound the two in order.

First consider the term $P_k(\mathrm{wt}(x)) - P_k(\mathrm{wt}(x+w))$ for some $x \in S_1$. Let $i = \mathrm{wt}(x)$. Then we have $\mathrm{wt}(x+w) \geq i - \delta n$ and so $P_k(\mathrm{wt}(x+w)) \leq P_k(i - \delta n)$ (using Proposition 3.3, Part (6)). Thus

$$\begin{aligned}
P_k(\mathrm{wt}(x)) - P_k(\mathrm{wt}(x+w)) &\geq P_k(i) - P_k(i - \delta n) \\
&= \sum_{j=i-\delta n+1}^{i} P_k(j) - P_k(j-1)
\end{aligned}$$

$$= \sum_{j=i-\delta n+1}^{i} -(P_{k-1}(j) + P_{k-1}(j-1)) \quad \text{(Using Proposition 3.3, Part (5))}$$
$$\geq (-2\delta n) \cdot P_{k-1}(i) \quad \text{(Using Proposition 3.3, Part (6))}$$

We thus get that $T_1 \geq (-2\delta n) \cdot \sum_{x \in S_1} P_{k-1}(\mathrm{wt}(x)) = (-2\delta n) \cdot \sum_{i=b+\delta n}^{\frac{5}{6}n} P_{k-1}(i) B_i^C$. Using the fact that $P_{k-1}(i) = P_{k-1}(n-i)$, we get that $\sum_{i=b+\delta n}^{\frac{5}{6}n} P_{k-1}(i) B_i^C = \sum_{i=\frac{1}{6}n}^{\frac{n}{2}-\delta n-n^{1-\gamma}} P_{k-1}(i) B_i^{C'}$, where $C' = C + 1^n$. This last term can be bounded using Lemmas 5.3 and 5.4 and we get $\sum_{i=b+\delta n}^{\frac{5}{6}n} P_{k-1}(i) B_i^C \leq ((5/6)^k + 12(2/3)^{k-2} + \theta(n^{-c})) \cdot P_{k-1}(0)$. Plugging it back into our bound for T_1, we get $T_1 \geq (-2\delta n) \cdot ((5/6)^k + 12(2/3)^{k-2} + \theta(n^{-c})) \cdot P_{k-1}(0)$. Using the crude bound $nP_{k-1}(0) \leq 2kP_k(0)$ (which holds for $n \geq 2k$), we get that $T_1 \geq -c_k \delta P_k(0)$ for $c_k = 4k((5/6)^k + 12(2/3)^{k-2}) + \theta(n^{-c})$. Picking k, n sufficiently large, we can ensure $c_k \leq \epsilon/2$ and so we have $T_k \geq -\epsilon/2 \cdot \delta \cdot P_k(0)$.

We now move to the second term $T_2 = \sum_{i=b}^{b+\delta n} P_k(i) B_i^C$. This part can be analysed as in the proof of Lemma 5.4 to get that $T_2 \geq -(12(4\delta)^{k-2} + \theta(n^{-c})) \cdot P_k(0)$. Again, by picking k, n to be sufficiently large, we can set $T_2 \geq -\epsilon/2 \cdot \delta \cdot P_k(0)$. Putting the two terms together, we get the lemma. ■

6.3 Testing general codes

Next we move to the task of building tests for general codes C based on tests for slightly-biased codes. We do so with a generic reduction that shows that if a code C is locally testable, then so is the code $C||v$ for any $v \in \{0,1\}^n$.

Definition 6.6 *Given linear test T for C, and $v \in \{0,1\}^n$ we define tests $T_v^{(1)}$ and $T_v^{(2)}$ for the code $C||v$ as follows:*

$T_v^{(1)}$: *Given oracle access to $w \in \{0,1\}^n$, accept if T^w accepts or T^{w+v} accepts.*

$T_v^{(2)}$: *Let $S \subseteq C^\perp$ be the "tests" of T. Fix a canonical $y_0 \in S$ such that $\langle y_0, v \rangle = 1$ if such a y_0 exists. Pick random $y \in S$ as drawn by the test T. If $\langle y, v \rangle = 0$ then accept iff $\langle y, w \rangle = 0$. If $\langle y, v \rangle = 1$ then accept iff $\langle y + y_0, w \rangle = 0$.*

We remark that even if T is a canonical test, $T_v^{(1)}$ and $T_v^{(2)}$ are not (or at least need not be) canonical. However they are both good tests, as shown below.

Lemma 6.7 *If C is (q, ϵ)-locally testable with the linear test T, then the following hold:*

1. *$T_v^{(1)}$ is a $(2q, \epsilon^2)$-test for $C||v$.*
2. *$T_v^{(2)}$ is a $(2q, \epsilon)$-test for $C||v$.*

In particular, if T is a strong test, then so are $T_v^{(1)}$ and $T_v^{(2)}$.

Proof: For Part (1), note first that $T_v^{(1)}$ does accept every word $w \in C||v$ with probability 1. On the other hand if $\delta(w, C||v) = \delta > 0$ then we have $\delta(w, C) \geq \delta$ and $\delta(w + v, C) \geq \delta$. Thus the probability that T would reject w is at least $\epsilon(\delta)$. Independently the probability that T would reject $w + v$ (on independent random coins) is also at least $\epsilon(\delta)$ and so the probability that $T_v^{(1)}$ rejects w is at least $\epsilon^2(\delta)$.

For Part (2), first note again that $T_v^{(2)}$ does accept every word $w \in C||v$ with probability 1. To see this consider a random choice $y \in S$ of the test T. For $w \in C$, we have $\langle y, w\rangle = \langle y_0, w\rangle = 0$ and so $\langle y + y_0, w\rangle = 0$ as well, so the test accepts in all cases. For $w \in C + v \Leftrightarrow w + v \in C$, if $\langle y, v\rangle = 0$ then $\langle y, w\rangle = \langle y, w + v\rangle = 0$ and the test accepts. On the other hand, if $\langle y, v\rangle = 1$, then $\langle y + y_0, w\rangle = \langle y + y_0, w + v\rangle + \langle y + y_0, v\rangle = 0 + \langle y + y_0, v\rangle = \langle y, v\rangle + \langle y_0, v\rangle = 1 + 1 = 0$ and so the test accepts in this case too. Thus we move to the soundness analysis of this test.

For this part, let p_0 denote the probability that the test T picks a y such that $\langle y, v\rangle = 0$. Fix $w \in \{0, 1\}^n$ such that $\delta(w, C||c) \geq \delta$. Let q_0 denote the probability that the test T picks y such that $\langle y, v\rangle = 0$ and $\langle y, w\rangle = 1$. Let q_1 denote the probability that the test T picks y such that $\langle y, v\rangle = 1$ and $\langle y, w\rangle = 0$. Let q_2 denote the probability that the test T picks y such that $\langle y, v\rangle = 1$ and $\langle y, w\rangle = 1$.

We claim first that the probability that the test $T_v^{(2)}$ rejects w is at least $q_0 + \min\{q_1, q_2\}$. To see this, consider the case when $\langle w, y_0\rangle = 0$. If the test picks y such that $\langle y, v\rangle = 0$ and $\langle y, w\rangle = 1$, then it surely rejects and this happens with probability at least q_0. On the other hand if $\langle y, v\rangle = \langle y, w\rangle = 1$ then $\langle y + y_0, w\rangle = \langle y, w\rangle + \langle y_0, w\rangle = 1 + 0 = 1$ and again the test rejects. This event is mutually exclusive of the previous one and happens with probability q_2. Thus in this case the test rejects with probability at least $q_0 + q_2$. Similarly, in the case $\langle w, y_0\rangle = 1$, we see that the test rejects with probability at least $q_0 + q_1$, yielding the claim.

But now we note that since $\delta(w, C) \geq \delta$, it must be that T rejects w with probability at least $\epsilon(\delta)$. But this probability is exactly $q_0 + q_2$. Similarly, since $\delta(w + v, C) \geq \delta$, we have that the probability that T rejects $w + v$ is at least $\epsilon(\delta)$. The probability that T rejects $w + v$ on the other hand is at least $q_0 + q_1$: The first quantity is the probability that $\langle y, v\rangle = 0$ and $\langle y, w + v\rangle = \langle y, w\rangle = 1$ while the second is the probability that $\langle y, v\rangle = 1$ and $\langle y, w + v\rangle = 1$ (where the $\langle y, w + v\rangle = 1$ iff $\langle y, v\rangle = 0$). We conclude thus that $q_0 + \min\{q_1, q_2\} \geq \epsilon(\delta)$. ■

The proof of Theorem 1.1 now follows easily.

Proof: [Theorem 1.1] Suppose C is $\frac{1}{3}$-biased, then the theorem follows from Lemma 6.4. Suppose C is not $\frac{1}{3}$-biased. Then let $v \in C$ be such that $\delta(0, v) = 1 - \tau > \frac{5}{6}$. Let C' a linear subcode of $C - \{v\}$ such that $C = C'||v$. Then C' is n^t-sparse, has distance $\frac{1}{2} - n^{-\gamma}$ (inheriting these properties from C). Most importantly, we note that C' is also moderately small-biased. To see this consider a codeword w of C' of weight $(1 - \epsilon) \cdot n$. Then $\delta(v, w) \leq \tau + \epsilon$. Since $\tau \leq 1/6$, it must be the case $\epsilon \geq \frac{1}{3} - n^{-\gamma} \geq \frac{1}{6}$ (for sufficiently large n). We conclude that C' has no codewords of weight more that $5/6 \cdot n$ making it a (1/3)-biased code. We can now apply Lemma 6.4 to conclude that C' is strongly locally testable, and thus by Lemma 6.7, $C = C'||v$ is also strongly locally testable. ■

7 On Sparse Random Codes

In this section we show that random linear codes with $O(n^t)$ codewords are self-correctible and testable with high probability, and that random linear encodings are locally decodable with high

probability.

In contrast we point out that codes with quasi-polynomially many codewords are not locally testable, by showing they have no small weight codewords in their dual.

The first part follows immediately from the following easy fact.

Proposition 7.1 *Let $E : \{0,1\}^{t \log n} \to \{0,1\}^n$ be a random linear map chosen by picking $A \in_U \{0,1\}^{n \times (t \log n)}$ and letting $E(m) = A \cdot m$. Let $C = \{E(m) | m \in \{0,1\}^{t \log n}\}$. Then with high probability C is $O(\log n / \sqrt{n})$-biased.*

We conclude with the following theorem (which follows easily from Theorem 1.2, Theorem 5.5, and Corollary 4.6).

Theorem 7.2 *For every $t < \infty$ there exists a constant $k = k_t < \infty$ such that a randomly chosen linear encoding (chosen by picking its $t \log n$ basis vectors uniformly from $\{0,1\}^n$ with replacement) is k-locally decodable, with probability tending to 1. Furthermore, the image of the encoding is k-locally testable and k-self-correctible, with probability tending to 1.*

Conversely we have:

Proposition 7.3 *For every t, if $C \subseteq \{0,1\}^n$ is a random linear code of size $2^{(\log n)^t}$, then $C^\perp$ has distance $\Omega((\log n)^{t-1})$ with high probability, and so can not be tested with $o((\log n)^{t-1})$ queries.*

Proof: Let $k = (\log n)^t$. Let $C \subseteq \{0,1\}^n$ be a randomly chosen linear code with 2^k codewords. Specifically C is chosen by picking a generator matrix $G \subseteq \{0,1\}^{k \times n}$ uniformly at random and letting $C = \{xG | x \in \{0,1\}^k\}$. The claim that $C^\perp$ has distance greater than $\ell = \Omega((\log n)^{t-1})$ is equivalent to saying that every subset of ℓ columns of G are linearly independent, or equivalently no subset of ℓ or fewer columns of G add up to zero. The probability that a specific (non-empty) subset sums to zero is at most 2^{-k}. By the union bound, the probability that there exists a subset of fewer than ℓ columns that sum to zero is at most $\sum_{i=0}^{\ell} \binom{n}{i} 2^{-k}$ which goes to zero if $\ell \ll \frac{k}{\log n}$. The proposition follows. ■

Acknowledgments

We would like to thank Oded Goldreich for suggesting that random codes may be testable using the techniques of [11], and for endless inspiration and enthusiasm. We are grateful to Simon Litsyn for many valuable discussions and encouragement. We would like to thank Swastik Kopparty for helpful discussions.

References

[1] László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy. Checking computations in polylogarithmic time. In *Proceedings of the 23rd ACM Symposium on the Theory of Computing*, pages 21–32, New York, 1991. ACM Press.

[2] Amos Beimel, Yuval Ishai, Eyal Kushilevitz, and Jean Francois Raymond. Breaking the $o(n^{1/(2k-1)})$ barrier for information-theoretic private information retrieval. In *Proceedings of the 43rd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002.

[3] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 1–10, New York, 2004. ACM Press.

[4] Eli Ben-Sasson, Oded Goldriech, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Short PCPs verifiable in polylogarithmic time. In *Proceedings of the Twelfth Annual IEEE Conference on Computational Complexity*, pages 120–134, June 12–15 2005.

[5] Eli Ben-Sasson and Madhu Sudan. Short PCPs with poly-log rate and query complexity. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 266–275, New York, 2005. ACM Press.

[6] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

[7] Gerard Cohen, Iiro Honkala, Simon Litsyn, and Antoine Lobstein. *Covering Codes*. North-Holland Mathematical Library, 54. North-Holland, Amsterdam, 1997.

[8] Irit Dinur. The PCP theorem by gap amplification. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 241–250, New York, 2006. ACM Press. Preliminary version appeared as an ECCC Technical Report TR05-046.

[9] Oded Goldreich and Madhu Sudan. Locally testable codes and PCPs of almost-linear length. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, Vancouver, Canada, 16-19 November 2002.

[10] Jonathan Katz and Luca Trevisan. On the efficiency of local decoding procedures for error-correcting codes. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 80–86, New York, NY, USA, 2000. ACM Press.

[11] T. Kaufman and S. Litsyn. Almost orthogonal linear codes are locally testable. In *Proceedings of the Forty-sixth Annual Symposium on Foundations of Computer Science*, pages 317–326, 2005.

[12] Ilia Krasikov and Simon Litsyn. *On Binary Krawtchouk Polynomials*, volume 56 of *DIMACS series in Discrete Mathematics and Theoretical Computer Science*, pages 199–212. American Mathematical Society, Providence, 2001.

[13] Joseph Naor and Moni Naor. Small-bias probability spaces: Efficient constructions and applications. *SIAM Journal on Computing*, 22(4):838–856, 1993.

[14] Sergey Yekhanin. Towards 3-query locally decodable codes of subexponential length. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing*, page (to appear). ACM Press, 2007.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de/