

# Locally Testable Codes Require Redundant Testers

Eli Ben-Sasson  
Dept. of Computer Science  
Technion, Haifa 32000, Israel  
Email: eli@cs.technion.ac.il

Venkatesan Guruswami  
Dept. of Computer Science & Engineering  
University of Washington,  
Seattle, WA 98195, USA

Tali Kaufman  
CSAIL, MIT  
Cambridge, MA 02139, USA  
Email: kaufmant@mit.edu

Madhu Sudan  
CSAIL, MIT  
Cambridge, MA 02139, USA  
Email: madhu@mit.edu

Michael Viderman  
Dept. of Computer Science  
Technion, Haifa 32000, Israel  
Email: viderman@cs.technion.ac.il

**Abstract**—Locally testable codes (LTCs) are error-correcting codes for which membership, in the code, of a given word can be tested by examining it in very few locations. Most known constructions of locally testable codes are linear codes, and give error-correcting codes whose duals have (superlinearly) *many* small weight codewords. Examining this feature appears to be one of the promising approaches to proving limitation results for (i.e., upper bounds on the rate of) LTCs.

Unfortunately till now it was not even known if LTCs need to be non-trivially redundant, i.e., need to have *one* linear dependency among the low-weight codewords in its dual. In this paper we give the first lower bound of this form, by showing that every positive rate constant query strong LTC must have linearly many redundant low-weight codewords in its dual. We actually prove the stronger claim that the *actual test itself* must use a linear number of redundant dual codewords (beyond the minimum number of basis elements required to characterize the code); in other words, non-redundant (in fact, low redundancy) local testing is impossible.

**Keywords**—linear codes; property testing; dual codes; LDPC codes; lower bounds.

## I. INTRODUCTION

In this work, we exhibit some *limitations* of locally testable linear codes. A linear code over a finite field  $\mathbb{F}$  is a linear subspace  $\mathcal{C} \subseteq \mathbb{F}^n$ . The dimension of  $\mathcal{C}$  is its dimension as a vector space, and its rate is the ratio of its dimension to  $n$ . The distance of  $\mathcal{C}$  is the minimal Hamming distance between two different codewords.

Research of first, fourth and fifth co-authors supported by grant number 2006104 by the US-Israel Binational Science Foundation. Work of the first and last co-authors supported by grant number 679/06 by the Israeli Science Foundation.

The work of the second author was done when on leave at the Computer Science Department, Carnegie Mellon University and during a visit to the Technion in 2008. Research supported in part by a Packard fellowship, and NSF grants CCF-0343672 and CCF-0835814.

One is typically interested in codes whose distance is a growing function of the block length  $n$ , ideally  $\Omega(n)$ . Such a code is *locally testable* if given a word  $x \in \mathbb{F}^n$  one can verify with good accuracy whether  $x \in \mathcal{C}$  by reading only a few (say a constant independent of  $n$ ) chosen symbols from  $x$ . More precisely such a code has a *tester*, which is a randomized algorithm with oracle access to the received word  $x$ . The tester reads at most  $q$  symbols from  $x$  and based on this local view decides if  $x \in \mathcal{C}$  or not. It should accept codewords with probability one, and reject words that are far (in Hamming distance) from the code with noticeable probability.

Locally Testable Codes (henceforth, LTCs) are the combinatorial core of PCP constructions. In recent years, starting with the work of Goldreich and Sudan [17], several surprising constructions of LTCs have been given (see [16] for an extensive survey of some of these constructions). The principal challenge is to understand the largest asymptotic rate possible for LTCs, and to construct LTCs approaching this limit. We now know constructions of LTCs of dimension  $n/\log^{O(1)} n$  which can be tested with only three queries [11], [13].

One of the outstanding open questions in the subject is whether there are asymptotically good LTCs, i.e., LTCs that have dimension  $\Omega(n)$  and distance  $\Omega(n)$ . Our understanding of the *limitations* of LTCs is, however, quite poor (in fact, practically non-existent), and approaches that may rule out the existence of asymptotically good LTCs have been elusive. Essentially the only negative results on LTCs concern binary codes testable with just 2-queries [8], [19] (which is a severe restric-

tion), random LDPC codes [10], and cyclic codes [4].<sup>1</sup> In fact, we cannot even rule out the existence of binary LTCs meeting the Gilbert-Varshamov bound (which is the best known rate for codes without any local testing restriction). So, for all we know, the strong testability requirement of LTCs may not “cost” anything extra over normal codes!

This work is a (modest) initial attempt at addressing our lack of knowledge concerning lower bound results for LTCs. For linear codes, one can assume without loss of generality [10] that the tester picks a low-weight dual codeword  $c^\perp$  from some distribution, and checks that the input  $x$  is orthogonal  $c^\perp$ . It is thus necessary that if  $C$  is a  $q$ -query LTC of dimension  $k$ , then its dual  $C^\perp$  has a basis of  $n - k$  codewords each of weight at most  $q$ .<sup>2</sup> All known constructions of LTCs in fact have duals which have super-linearly many low-weight dual codewords. In other words, there must be a substantial number of linear dependencies amongst the low-weight dual codewords. Examining whether this feature is necessary might be one of the promising approaches to proving limitations (i.e., upper bounds on the rate) of LTCs, as it imposes strong constraints on the dual code.<sup>3</sup> Nevertheless, till now it was not even known if the dual of a LTC has to be non-trivially redundant, i.e., if it must have at least *one* linear dependency among its low-weight codewords.

In this work, we give the first lower bound of this form, by showing that every positive rate constant query LTC must have  $\Omega(n)$  redundant low-weight codewords. The result is actually stronger — it shows that the *actual test itself* must use  $\Omega(n)$  extra redundant dual codewords (beyond the minimum  $n - k$  basis elements). In other words, *non-redundant testing is impossible*. While this might sound like an intuitively obvious statement, we remark that even for Hadamard codes (whose dual has  $\Theta(n^2)$  weight 3 codewords), a non-redundant test consisting of a basis of weight 3 dual codewords was not ruled out prior to our work. Also, without the restriction on number of queries, *every* code

<sup>1</sup>The last result rules out asymptotically good *cyclic* LTCs; the existence of asymptotically good cyclic codes has been a longstanding open problem, and the result shows the “intersection” of these questions concerning LTCs and cyclic codes has a negative answer.

<sup>2</sup>To be precise, only when  $C$  is a *strong* LTC, as per Definition 2, need  $C^\perp$  be spanned by words of weight  $q$ . Non-strong LTCs have the property that the set of low-weight words in the dual code must span a large dimensional subspace of  $C^\perp$  (see Proposition 15 for an exact statement).

<sup>3</sup>We remark that information on the dual weight distribution is useful, for example, in the linear programming bounds on the rate vs. distance trade-off of a linear code. For LDPC codes whose dual has a low weight basis, stronger upper bounds on distance are known compared to general linear codes of the same rate [6].

does admit a basis tester (which makes at most  $k + 1$  queries). We also note that a known upper bound [5, Proposition 11.2] shows that  $O(n)$  redundancy suffices for testing. [5] prove this in the context of PCPs, but the technique extends to LTCs as well. For completeness, in Section VII, we include a proof showing that for *every*  $q$ -query LTC, there is a  $O(q)$ -query tester that picks a test uniformly from at most  $3(n - k) = O(n)$  dual codewords. The quantity  $n - k$  (as opposed to  $n$ ) is significant in that this is the dimension of the dual code, and our lower bound shows that every tester (for any code) must have a support of size at least  $n - k$ .

## II. DEFINING THE REDUNDANCY OF A TESTER

*Preliminary notation:* Throughout this paper  $\mathbb{F}$  is a finite field,  $[n]$  denotes the set  $\{1, \dots, n\}$  and  $\mathbb{F}^n$  denotes  $\mathbb{F}^{[n]}$ . For  $w = \langle w_1, \dots, w_n \rangle \in \mathbb{F}^n$  let  $\text{supp}(w) = \{i \mid w_i \neq 0\}$  and  $\text{wt}(w) = |w| = |\text{supp}(w)|$ .

We define the *distance* between two words  $x, y \in \mathbb{F}^n$  to be  $\Delta(x, y) = |\{i \mid x_i \neq y_i\}|$  and the relative distance to be  $\delta(x, y) = \frac{\Delta(x, y)}{n}$ .

We use the standard notation for describing linear error correcting codes and point out that all codes discussed in this paper are linear. A  $[n, k, d]_{\mathbb{F}}$ -code is a  $k$ -dimensional subspace  $C \subseteq \mathbb{F}^n$  of distance  $d$ , defined next. The relative distance of  $C$  is denoted  $\delta(C)$  and defined to be the minimal value of  $\delta(x, y)$  for two distinct codewords  $x, y \in C$ . The distance of  $C$  is  $\Delta(C) = \delta(C) \cdot n$ . Let  $\delta(x, C) = \min_{y \in C} \{\delta(x, y)\}$  denote the relative distance of  $x$  from the code  $C$ . We say that  $x$  is  $\alpha$ -far from  $C$  if  $\delta(x, C) \geq \alpha$  and otherwise we say  $x$  is  $\alpha$ -close to  $C$ . The inner-product between two vectors  $u$  and  $v$  in  $\mathbb{F}^n$  is  $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$ .

For a linear code  $C$  let  $C^\perp$  denote its dual code, i.e.,  $C^\perp = \{u \in \mathbb{F}^n \mid \forall c \in C : \langle u, c \rangle = 0\}$  and recall  $\dim(C^\perp) = n - \dim(C)$ . Let  $C_{<t}^\perp = \{u \in C^\perp \mid |u| < t\}$  and  $C_{\leq t}^\perp = \{u \in C^\perp \mid |u| \leq t\}$ .

**Definition 1 (Tester).** Suppose  $C$  is a  $[n, k, d]_{\mathbb{F}}$ -code. A  $q$ -query test for  $C$  is an element  $u \in C_{\leq q}^\perp$  and a  $q$ -query tester  $T$  for  $C$  is defined by a distribution  $p$  over  $q$ -query tests. When  $C$  is clear from context we omit reference to it. The *support* of  $T$ , denoted  $S = S_T$ , is the support of  $p$ , i.e., the set  $S = S_T = \{u \in C_{\leq q}^\perp \mid p(u) > 0\}$ . When  $p$  is uniform over a subset of  $C_{\leq q}^\perp$  we say the tester is *uniform* and may identify the tester with  $S$ .

Invoking the tester  $T$  on a word  $w \in \mathbb{F}^n$  is done by sampling a test  $u \in S_T$  according to the distribution  $p$  and outputting **accept** if  $\langle u, w \rangle = 0$ , in which case we say that  $u$  (and  $T$ ) **accept**  $w$ , denoted  $T[w] = \text{accept}$ , and outputting **reject**, denoted  $T[w] = \text{reject}$ ,

if  $\langle u, w \rangle \neq 0$ . Clearly any such tester always accepts  $w \in \mathcal{C}$ .

- A  $(q, \rho')$ -strong tester is a  $q$ -query tester  $T$  satisfying for all  $w \in \mathbb{F}^n$

$$\Pr[T[w] = \text{reject}] \geq \rho' \cdot \delta(w, \mathcal{C}).$$

- A  $(q, \varepsilon, \rho)$ -tester is a  $q$ -query tester  $T$  satisfying for all  $w \in \mathbb{F}^n$  that is  $\varepsilon$ -far from  $\mathcal{C}$

$$\Pr[T[w] = \text{reject}] \geq \rho.$$

The probability in both equations above is according to the distribution  $p$  associated with  $T$ .

**Definition 2** (Locally Testable Code (LTC) [17]). A  $[n, k, d]_{\mathbb{F}}$ -code  $\mathcal{C}$  is said to be a  $(q, \rho')$ -strong locally testable code if it has a  $(q, \rho')$ -strong tester, and  $\mathcal{C}$  is a  $(q, \varepsilon, \rho)$ -locally testable code if it has a  $(q, \varepsilon, \rho)$ -tester. The parameter  $\rho$  is known as the *soundness* of  $T$  and  $\varepsilon$  is its *distance parameter*.

Note that a  $(q, \rho')$ -strong LTC is also a  $(q, \varepsilon, \rho' \cdot \varepsilon)$  LTC for every  $\varepsilon > 0$ . Moreover, if  $T$  is a  $(q, \rho' > 0)$ -strong tester for a  $[n, k, d]_{\mathbb{F}}$ -code then, letting  $S_T$  denote the support of  $T$ , we have  $\dim(S_T) = \dim(\mathcal{C}^\perp) = n - k$ .

*Remarks on definitions of testers:* Our definition of a tester, and an LTC is somewhat different from previous definitions (notably [10] and [17]). We clarify the differences here.

We start with Definition 2. The definition of strong LTCs we use is the same as that in [17]. The weak notion is weaker than their definition of a weak tester (which simply allowed the rejection probability of a weak tester to be smaller by a  $o(1)$  additive amount compared to the strong case). Our definition on the other hand only requires rejection probability to be positive when the word is very far (constant relative distance) from the code. Since our goal is to prove “impossibility” results, doing so with weaker definitions makes our result even stronger.

We now discuss Definition 1. For linear LTCs it was shown in [10] (see also references therein) that the tester might as well pick a collection of low-weight dual codewords and verify that the given word  $w$  is orthogonal to all of them. On the other hand, our definition (Definition 1) requires the tester to pick only one dual codeword and test orthogonality to it. Our definition is more convenient to us when defining and analyzing the *redundancy* of tests (defined below). We first note that our restricted forms of tests may only alter the soundness of the test by a constant factor.

For this we recall the assertion from [10] who showed

that without loss of generality a  $q$ -query “standard” tester for a  $[n, k, d]_{\mathbb{F}}$ -code is defined by a distribution over subsets  $I \subseteq [n], |I| \leq q$ . The test associated with  $I$  accepts a word  $w$  if and only if  $\langle w, u \rangle = 0$  for all  $u \in \mathcal{C}^\perp$  such that  $\text{supp}(u) \subseteq I$ . (The soundness and distance parameters of a “standard” tester are defined as in Definition 1.) To convert this “standard” tester to one that only tests one dual codeword, consider a tester that, given  $I$ , samples uniformly from the set  $U_I = \{u \in \mathcal{C}^\perp \mid \text{supp}(u) \subseteq I\}$  and accepts iff  $\langle u, w \rangle = 0$ . This resulting tester conforms to our Definition 1. Furthermore, if the soundness of the “standard” tester is  $\rho$  then the soundness of the tester that samples uniformly from  $U_I$  is at least  $\frac{|\mathbb{F}|-1}{|I|} \rho \geq \frac{1}{2} \rho$ . To see this, notice that  $U_I$  forms a linear space over  $\mathbb{F}$ . And the set  $\{u \in U_I \mid \langle u, w \rangle = 0\}$  is a linear subspace of  $U_I$ . Thus, whenever  $w$  is rejected by some  $u \in U_I$  we actually know that  $w$  is rejected by at least a fraction  $\frac{|\mathbb{F}|-1}{|I|}$  of  $U_I$  because the set of rejecting words is the complement of a subspace of  $U_I$ . Hence, using our definition of a tester is equivalent to the most general definition of a tester, up to a constant loss in the soundness parameter.

**Definition 3** (Basis Tester and Tester redundancy). Suppose  $\mathcal{C}$  is a  $[n, k, d]_{\mathbb{F}}$ -code. A  $q$ -query tester  $T$  for  $\mathcal{C}$  is said to be a *basis tester* if its support  $S_T$  forms a basis for  $\mathcal{C}^\perp$ , i.e.,  $S_T \subseteq \mathcal{C}_{\leq q}^\perp$  is linearly independent and of size  $\dim(\mathcal{C}^\perp) = n - k$ . In case  $S_T$  spans  $\mathcal{C}^\perp$  but has size larger than  $\dim(\mathcal{C}^\perp)$  we define the *redundancy* of  $T$  to be  $|S_T| - \dim(\mathcal{C}^\perp)$ . (Notice a basis tester has redundancy 0.)

### III. MAIN RESULTS

**Theorem 4** (Main Theorem). *If a  $[n, k, d]_{\mathbb{F}}$ -code  $\mathcal{C}$  has a  $(q, \frac{\delta(\mathcal{C})}{3}, \rho)$ -basis tester then*

$$\rho \leq \frac{q}{k}.$$

**Remark 5.** The main theorem holds even for a basis tester that has only *expected* query complexity  $\leq q$  (and all other parameters are as in the statement of the theorem). We say a basis tester has *expected* query complexity at most  $q$  if  $\mathbb{E}_u[|u|] = q$  where the expectation is taken with respect to the probability associated with the tester. Additionally, the theorem holds even for a tester whose support is *strictly contained* in a basis for  $\mathcal{C}^\perp$ .

The first corollary of our main theorem says that  $\Omega(n)$  redundancy is necessary for uniform testing of all codes that have nontrivial (i.e., super-constant) size.

**Corollary 6** (Uniform testers for LTCs with super constant size require linear redundancy). *Let  $\mathcal{C}$  be a  $[n, k, d]$*

code that is  $(q, \frac{1}{3}\delta(\mathcal{C}), \rho)$ -locally testable by a uniform tester using a set  $S \subseteq \mathcal{C}_{\leq q}^\perp$  with  $\dim(S) = \dim(\mathcal{C}^\perp)$ . Then

$$|S| \geq \left( \frac{1 - q/k}{1 - \rho} \right) \cdot \dim(\mathcal{C}^\perp).$$

In words,  $S$  has redundancy at least  $\frac{\rho - q/k}{1 - \rho} \cdot \dim(\mathcal{C}^\perp)$ .

For instance, if  $k = \dim(\mathcal{C}) = \omega(1)$  and  $\rho, q$  are constants then the previous corollary says that a uniform tester for  $\mathcal{C}$  requires a linear amount ( $\Omega(n)$ ) of redundancy.

Our second corollary shows that non-trivial redundancy is necessary for general (i.e., for nonuniform) testing.

**Corollary 7** (Testers for LTCs with constant rate require linear redundancy). *Let  $\mathcal{C}$  be a  $[n, k, d]$  code that is  $(q, \frac{1}{3}\delta(\mathcal{C}), \rho)$ -locally testable by a tester that is distributed over a set  $S \subseteq \mathcal{C}_{\leq q}^\perp$  of dimension  $\dim(S) = \dim(\mathcal{C}^\perp)$ . Then*

$$|S| \geq \dim(\mathcal{C}^\perp) + \frac{\rho k}{q} - 1.$$

In words,  $S$  has redundancy at least  $\frac{\rho k}{q} - 1$ .

For instance, if  $k = \Theta(n)$  and  $\rho, q$  are constants (i.e., when  $\mathcal{C}$  comes from an asymptotically good family of error correcting codes) then, once again, a linear amount of redundancy is required by any constant-query tester for  $\mathcal{C}$ . For the state of the art LTCs [11], [13], [22]  $k = \Theta(n/\text{poly}(\log n))$  and our result implies that  $\Theta(n/\text{poly}(\log n))$  redundancy is necessary in such cases.

Finally, we show that our main theorem is almost tight in two respects. In section VI we show that there do exist codes of constant size that can be strongly tested by a uniform basis tester and that every code can be strongly tested by a uniform basis tester that has large query complexity. We conclude by showing in Section VII that if  $\mathcal{C} \subseteq F_2^n$  is a  $(q, \varepsilon, \rho)$ -LTC then it has a  $(\frac{10q}{\rho}, \varepsilon, \frac{1}{100})$ -tester that is uniform over a multiset  $S$  with a small (linear) amount of redundancy, i.e., with  $|S| \leq 3 \dim(\mathcal{C}^\perp)$  and  $\dim(S) \geq \dim(\mathcal{C}^\perp) - 3\varepsilon n$ .

#### IV. PROOF OF MAIN THEOREM 4

##### A. Overview

To explain what goes on in the proof we focus on a relatively simple case. We say that a tester is *smooth* if it has the property that every bit of the input word  $w$  is queried by it with equal probability. Let us sketch how to prove a linear lower bound on the query complexity  $q$  of a  $(q, \rho)$ -strong smooth and uniform basis tester for

a  $[n, k = \kappa n, d = \delta n]_{\mathbb{F}_2}$ -code  $\mathcal{C}$  over the two-element field  $\mathbb{F}_2$ . Namely, we will show  $q = \Omega(n)$ .

Let  $B = \{u_1, \dots, u_{n-k}\}$  be the set of tests selected (uniformly) by our smooth basis tester  $T$ . By assumption  $B$  is a basis for  $\mathcal{C}^\perp$  and contains words of size at most  $q$ .

The main idea implemented in the proof is to build a special basis for  $\mathbb{F}_2^n$  using the code  $\mathcal{C}$  and the basis  $B$ . Specifically we define a set  $V = \{v_1, \dots, v_{n-k}\}$  such that for every word  $w \in \mathbb{F}_2^n$  we can find a codeword  $c_w \in \mathcal{C}$  and a set  $V_w \subseteq V$  such that  $w = c_w + \sum_{v \in V_w} v$ . (Specifically, we build such a set  $V$  by letting  $v_j \in \mathcal{C}^\perp$  such that  $v_j$  has inner product zero with  $u_i$  for every  $i \neq j$  and inner product one with  $u_j$ .)

We note that in this basis, the rejection probability of the basis tester based on  $B$  is straightforward to compute. A word  $w$  is rejected with probability exactly  $|V_w|/|V|$ . (This follows from the fact that  $u_i$  rejects  $w$  iff  $v_i \in V_w$ .)

Since this applies also to the elements  $v_i \in V$  also, we conclude they have small weight. Specifically, using the assumption that  $B$  is a  $(q, \rho)$ -strong tester we conclude

$$\begin{aligned} \rho \cdot \frac{|v_i|}{n} &= \rho \delta(v_i, \mathcal{C}) \leq \Pr[T[v_i] = \text{reject}] \\ &= \frac{|\{v_i\}|}{|V|} \leq \frac{1}{(1 - \kappa)n} \end{aligned}$$

which gives  $|v_i| \leq \frac{1}{\rho(1 - \kappa)} = O(1)$ .

The non-trivial step now is to consider the probability of rejecting some *low-weight* words. Specifically we consider the probability of rejecting the “unit” vector  $e_i$  in the standard basis. I.e.,  $e_i = 0^{i-1}10^{n-i}$ . On the one hand, smoothness implies this word can not be rejected with high-probability if the query complexity is low (since its weight is so low). On the other hand, we note that for some  $i$ , the set  $V_{e_i}$  has to be large and so it must be rejected with high probability. This leads to a contradiction to the assumption that the query complexity is low. We give more details below.

Note that there must exist a vector  $e_i$  whose representation is

$$e_i = c_{e_i} + \sum_{v_j \in V_{e_i}} v_j$$

where  $c_{e_i}$  is a *nonzero* codeword. This is because  $e_1, \dots, e_n$  are linearly independent, so they cannot all belong to  $\text{span}(V)$  which is a  $(n - k)$ -dimensional space. The crucial observation is that  $|V_{e_i}|$  must be large. This is because  $|v_j| \leq \frac{1}{\rho(1 - \kappa)}$  and  $|c_{e_i}| \geq \delta n$  so  $|V_{e_i}| \geq \frac{\delta}{\rho(1 - \kappa)} n$ . This implies that  $e_i$  is rejected with



probability

$$\frac{|V_{e_i}|}{|V|} \geq \frac{\frac{\delta}{\rho(1-\kappa)}n}{(1-\kappa)n} = \frac{\delta}{\rho}.$$

On the other hand, the assumption of smoothness implies rejection probability of  $e_i$  is precisely the probability of querying the  $i$ th coordinate which is  $\frac{q}{n-k} = \frac{q}{(1-\kappa)n}$ . We conclude

$$\frac{\delta}{\rho} \leq \frac{|V_{e_i}|}{|V|} = \Pr[T[e_i] = \text{reject}] = \frac{q}{(1-\kappa)n}$$

which gives  $q \geq \frac{\delta(1-\kappa)}{\rho}n = \Omega(n)$  as claimed.

Our proof of Theorem 4 follows the outline laid above. The two noticeable differences are that the tester need not be smooth, nor uniform, and the field size may be greater than 2. But the overall picture is roughly the same. Now for the details.

#### B. The $(\mathcal{C}, V)$ -representation of words in $\mathbb{F}^n$

Let  $B = \{u_1, \dots, u_{n-k}\} \subseteq \mathcal{C}_{\leq q}^\perp$  be a basis for  $\mathcal{C}^\perp$ .

The first part of our proof shows that every word in  $\mathbb{F}^n$  can be represented uniquely as the sum of a codeword in  $\mathcal{C}$  and a subset of a set of vectors  $V = \{v_1, \dots, v_{n-k}\}$  where the rejection probability of  $w$  is related to its representation structure. We start by defining  $V$ .

**Definition 8.** For  $i \in [n-k]$  let  $v_i$  to be a word of minimal weight that satisfies

$$\langle v_i, u_j \rangle = \begin{cases} 1 & i = j \\ 0 & j \in [n-k] \setminus \{i\} \end{cases} \quad (1)$$

and let  $V = \{v_1, \dots, v_{n-k}\}$ .

**Proposition 9.** For all  $v_i \in V$  we have  $\frac{\text{wt}(v_i)}{n} = \delta(v_i, \mathcal{C})$ .

*Proof:* We have  $\delta(v_i, \mathcal{C}) \leq \frac{\text{wt}(v_i)}{n}$  because  $\delta(v_i, 0^n) = \frac{\text{wt}(v_i)}{n}$  and  $0^n \in \mathcal{C}$ . On the other hand, for every  $c \in \mathcal{C}$  we must have  $\delta(v_i, c) \geq \frac{\text{wt}(v_i)}{n}$  because if  $\delta(v_i, c) < \frac{\text{wt}(v_i)}{n}$  then setting  $v'_i = v_i - c$  we have  $\text{wt}(v'_i) < \text{wt}(v_i)$  but  $v'_i$  satisfies (1) (with respect to index  $i$ ), thus contradicting the minimal weight of  $v_i$ . ■

The following claim states that  $\mathbb{F}^n$  is the direct sum of the code  $\mathcal{C}$  and  $\text{span}(V)$ .

**Claim 10.**  $\dim(\text{span}(\mathcal{C} \cup V)) = n$  and  $\dim(V) = n-k$ .

*Proof:* Let  $S = \mathcal{C} \cup V$ . To prove both equalities stated in our claim it is sufficient to show that  $S^\perp = \{0^n\}$ , i.e., that  $\dim(S^\perp) = 0$ , because  $\dim(\mathcal{C}) = k$  and  $|V| = n-k$ . Assume by way of contradiction that  $u \in S^\perp$  is nonzero. Then in particular  $u \in \mathcal{C}^\perp$  because  $\mathcal{C} \subseteq$

$S$  which implies  $\mathcal{C}^\perp \supseteq S^\perp$ . Thus,  $u$  is a nonzero linear combination of vectors from  $B$  because  $B$  is a basis for  $\mathcal{C}^\perp$ . Suppose  $u_i$  appears in the representation of  $u$  under  $B$ . Then from (1) we conclude  $\langle u, v_i \rangle \neq 0$  which implies  $u \notin V^\perp$  which gives  $u \notin S^\perp$ , contradicting the assumption  $u \in S^\perp$ . So  $\dim(S^\perp) = 0$  and this completes our proof. ■

Claim 10 shows that every  $w \in \mathbb{F}^n$  has a unique representation as a sum of a single element from  $\mathcal{C}$ , denoted  $c(w)$ , and a linear combination of  $v_j$ 's, denoted  $v(w)$ . We say  $(c(w), v(w))$  is the  $(\mathcal{C}, V)$ -representation of  $w$ . We denote by  $\Gamma(w) \subseteq [n-k]$  the set of indices  $(j)$  of  $v_j$ 's participating in  $v(w)$ . Formally, if  $v(w) = \sum_{j=1}^{n-k} \alpha_j v_j$  then

$$\Gamma(w) = \{j \mid \alpha_j \neq 0\}.$$

The next claim relates the rejection probability of  $w$  by our basis tester to the structure of  $v(w)$ . For  $i \in [n-k]$  let  $p(i) = p(u_i)$  denote the probability of  $u_i$  under the distribution associated with our basis tester. For  $I \subseteq [n-k]$  the set of indices of  $B' \subseteq B$  let  $p(I) = p(B') = \sum_{i \in I} p(i) = \sum_{u_i \in B'} p(u_i)$ .

**Claim 11** (Rejection probability is related to  $(\mathcal{C}, V)$ -representation structure). For all  $w \in \mathbb{F}^n$  we have

$$\Gamma(w) = \{j \in [n-k] \mid \langle u_j, w \rangle \neq 0\}. \quad (2)$$

Consequently, we have

$$\Pr[T[w] = \text{reject}] = p(\Gamma(w)).$$

*Proof:* Consider the  $(\mathcal{C}, V)$ -representation of  $w$ :

$$w = c(w) + \sum_{j \in \Gamma(w)} \alpha_j v_j, \text{ where } \alpha_j \neq 0.$$

By assumption for all  $u_i \in B$  we have  $\langle u_i, c(w) \rangle = 0$  and by (1) we have  $\langle u_i, v(w) \rangle \neq 0$  if and only if  $i \in \Gamma(w)$ . This implies (2). The consequence follows because, by definition, the probability of rejecting  $w$  is the probability of the event  $\langle u_i, w \rangle \neq 0$  where  $u_i$  is selected from  $B$  with probability  $p(i)$ . This completes the proof. ■

#### C. Main Lemma and Proof of Main Theorem 4

The following lemma is the main part of our proof. Assuming it we can complete the proof of Theorem 4 and the proof of the lemma comes after the proof of the theorem. In what follows the *singleton vector*  $e_i$  is the characteristic vector of the singleton set  $\{i\} \subset [n]$ .

**Lemma 12** (Main Lemma). There exists  $i \in [n]$  such that following two conditions hold:

- The probability that the  $i$ th bit is queried by the basis tester is at most  $q/k$ . Formally:  $p(\Gamma(e_i)) \leq q/k$ .
- Letting  $c(e_i) + v(e_i)$  denote the  $(\mathcal{C}, V)$ -representation of  $e_i$ , we have  $|v(e_i)| \geq d - 1$ .

*Proof of Theorem 4:* We show that Lemma 12 implies the existence of  $w \in \text{span}(\{v_j \mid j \in \Gamma(e_i)\})$  that is  $\frac{d}{3n}$ -far from  $\mathcal{C}$ . This implies

$$\rho \leq \Pr[T[w] = \text{reject}] \leq p(\Gamma(e_i)) \leq \frac{q}{k}$$

which implies the theorem.

First, observe that Claim 11 implies

$$\Gamma(e_i) = \{j \mid \text{supp}(u_j) \ni i\}. \quad (3)$$

Now, if  $|v_j| \geq \frac{1}{3}d$  for some  $j \in \Gamma(e_i)$  then setting  $w = v_j$  completes the proof. This is because Proposition 9 implies that  $v_j$  is  $\frac{d}{3n}$ -far from  $\mathcal{C}$  and Claim 11 implies it is rejected with probability  $p(j) \leq p(\Gamma(e_i)) \leq q/k$ .

From here on we assume  $|v_j| < \frac{1}{3}d$  for all  $j \in \Gamma(e_i)$ . Let  $t = |\Gamma(e_i)|$  and assume wlog  $\Gamma(e_i) = [t]$ . Denote the  $(\mathcal{C}, V)$ -representation of  $e_i$  by  $c(e_i) + \sum_{j=1}^t \alpha_j v_j$  where  $\alpha_j \neq 0$ . Let  $w_\ell = \sum_{j=1}^\ell \alpha_j v_j$ . We know the following:

- $|w_1| < \frac{1}{3}d$ .
- $|w_t| = |v(e_i)| \geq d - 1$  by the second bullet in Lemma 12.
- $|w_{\ell+1}| \leq |w_\ell| + |v_{\ell+1}| < |w_\ell| + \frac{1}{3}d$  for all  $1 \leq \ell < t$ , by the triangle inequality.

This implies the existence of some  $\ell \in [t]$  such that  $\frac{1}{3}d < |w_\ell| \leq \frac{2d}{3}$  and notice  $w = w_\ell$  is  $\frac{1}{3}d$ -far from  $\mathcal{C}$ . Claim 11 implies that  $w$  is rejected with probability

$$\Pr[T[w] = \text{reject}] = \sum_{j=1}^\ell p(j) \leq p(\Gamma(e_i)) \leq q/k$$

and this completes the proof of the Main Theorem 4. ■

*Proof of Lemma 12:* We start by showing that there exist  $k$  distinct singleton vectors, denoted without loss of generality  $e_1, \dots, e_k$ , such that  $c(e_1), \dots, c(e_k)$  are linearly independent, hence distinct and nonzero.

Since every word in  $\mathbb{F}^n$  has a unique  $(\mathcal{C}, V)$ -representation we get  $e_i \in \{c(e_i) + v \mid v \in \text{span}(V)\}$ . This implies

$$\{e_1, \dots, e_n\} \subseteq \text{span}(\{c(e_1), \dots, c(e_n)\} \cup V).$$

Counting dimensions, we have

$$\begin{aligned} n &= \dim(\text{span}(\{e_1, \dots, e_n\})) \\ &\leq \dim(\text{span}(\{c(e_1), \dots, c(e_n)\} \cup V)) \\ &\leq \dim(\text{span}(\{c(e_1), \dots, c(e_n)\})) + \dim(\text{span}(V)). \end{aligned}$$

By Claim 10 we have  $\dim(\text{span}(V)) = n - k$  we conclude that (without loss of generality)  $c(e_1), \dots, c(e_k)$  are linearly independent, as claimed.

Next, we argue that for  $i \in [k]$  we have  $|v(e_i)| \geq d - 1$ . This is because  $e_i = c(e_i) + v(e_i)$  and  $|e_i| = 1$  and  $|c(e_i)| \geq d$  because  $c(e_i)$  is a nonzero word in a linear code with minimal distance  $d$ .

So far we have shown that every  $v(e_i), i \in [k]$  satisfies the second bullet of the Main Lemma 12. To complete the proof all that remains is to show that for some  $i \in [k]$  we have  $p(\Gamma(e_i)) \leq q/k$ , because (3) shows that  $p(\Gamma(e_i))$  is the probability that our tester queries the  $i$ th bit. Using the assumption that  $T$  is a  $q$ -tester, i.e., that  $|u_j| \leq q$ , sum over all  $j \in [n - k]$  and get

$$\begin{aligned} \sum_{j=1}^k p(\Gamma(e_j)) &\leq \sum_{j=1}^n p(\Gamma(e_j)) = \sum_{j=1}^{n-k} p(u_j) \cdot |u_j| \\ &= \mathbf{E}_{u_j \sim p} [|u_j|] \leq q. \end{aligned}$$

We conclude that there must exist  $i \in [k]$  such that  $p(\Gamma(e_i)) \leq q/k$  and this completes the proof of Lemma 12. ■

## V. TESTING OF NONTRIVIAL CODES REQUIRES REDUNDANCY

In this section we prove our two main corollaries — 6 and 7.

*Proof of Corollary 6:* By assumption  $\dim(S) = \dim(\mathcal{C}^\perp) = n - k$ . Partition  $S$  into  $B \cup S'$  where  $B$  is a basis for  $\mathcal{C}^\perp$ ,  $|B| = n - k$  and  $S' = S \setminus B$  is the set of redundant tests. We bound the size of  $S'$  from below.

Consider a basis tester defined by  $B$ . By Theorem 4 this tester is not very sound, i.e., there exists a word  $w \in \mathbb{F}^n$  that is  $(\frac{1}{3}\delta(\mathcal{C}))$ -far from  $\mathcal{C}$  and is rejected by at most a fraction  $\rho_B \leq \frac{q(n-k)}{k}$  of the constraints in  $B$ . The overall number of constraints rejecting  $w$  is at least  $\rho|S| = \rho(|B| + |S'|)$  because  $S$  is a uniform tester for  $\mathcal{C}$  and  $w$  is far from  $\mathcal{C}$ . Taking the most extreme case that all words in  $S'$  reject  $w$  we get

$$\begin{aligned} \rho((n - k) + |S'|) &\leq |\{u \in S \mid \langle u, w \rangle \neq 0\}| \\ &\leq \frac{q}{k}(n - k) + |S'| \end{aligned}$$

which implies

$$|S'| \geq \frac{\rho - (q/k)}{1 - \rho} \cdot (n - k)$$

and this completes the proof of Corollary 6.  $\blacksquare$

*Proof of Corollary 7:* The high level idea is to partition  $S$  into a basis  $B$  for  $\mathcal{C}^\perp$  and a set of redundant tests  $S'$  such that, roughly speaking, the probability of sampling from  $B$ , according to the distribution  $p$  associated with  $T$ , is large. Then we continue as in the proof of Corollary 6.

To construct the said partition start with an arbitrary partition  $S = B \cup S'$  with  $B$  a basis for  $\mathcal{C}^\perp$ . Iteratively modify the partition as follows. If there exists  $u \in S'$  represented in the basis  $B$  as  $\sum_{b \in B} \alpha_b b$  and  $p(b) < p(u)$  for some  $b \in B$  with  $\alpha_b \neq 0$ , then replace  $b$  with  $u$ , i.e., set  $B$  to be  $(B \cup \{u\}) \setminus \{b\}$  and  $S'$  to be  $(S' \cup \{b\}) \setminus \{u\}$ . Repeat the process until no such  $u \in S'$  exists. Notice the process must terminate because  $\sum_{b \in B} p(b)$  is bounded by 1 and there exists  $\gamma > 0$  such that with each iteration this sum increases by at least  $\gamma$ .

At the end of the process we have partitioned  $S$  into a basis  $B$  for  $\mathcal{C}^\perp$  and a redundant set  $S'$  with the following property that will be crucial to our proof. For  $u \in S'$ , letting  $B(u)$  denote the minimal subset of  $B$  required to represent  $u$ , i.e.,  $B(u)$  satisfies

$$u = \sum_{b \in B(u)} \alpha_b b \text{ where } \alpha_b \neq 0,$$

then  $p(u) \leq p(b)$  for all  $b \in B(u)$ .

We continue with our proof. Consider the basis tester  $T'$  defined by taking the conditional distribution of our tester on  $B$  and let  $p'$  denote the conditional distribution on  $B$ , noticing  $p'(b) \geq p(b)$  for all  $b \in B$ . By theorem 4 there exists  $w$  that is  $\frac{1}{3}\delta(\mathcal{C})$ -far from  $\mathcal{C}$  and is rejected by  $T'$  with probability at most  $q/k$ . Let  $B' \subseteq B$  be the set of tests that reject  $w$  and notice  $p(B') \leq p'(B') \leq q/k$ .

Consider a word  $u \in S'$  that rejects  $w$  and represent  $u$  as a linear combination of elements of  $B(u) \subseteq B$ . Note that if the test  $u$  rejects  $w$  then there must be some  $b$  in  $B(u)$  that also rejects  $w$  (and hence belongs to  $B'$ ). By the special properties of our partition which were discussed in the previous paragraph we have

$$p(u) \leq p(b) \leq p(B') \leq p'(B') \leq q/k.$$

Thus, every test that rejects  $w$  from  $S'$  has probability at most  $q/k$  of being performed and furthermore, the probability of rejecting  $w$  using an element of  $B$  is at

most  $q/k$  as well. Summing up, we get

$$\rho \leq \Pr[T[w] = \text{reject}] \leq q/k + |S'| \cdot q/k$$

which after rearranging the terms give  $|S'| \geq \frac{\rho k}{q} - 1$  as claimed.  $\blacksquare$

## VI. TIGHTNESS OF MAIN THEOREM 4

In this section we argue that the bound ( $k \leq \frac{q}{\rho}$ ) obtained in Theorem 4 is close to tight. In Proposition 13 we show that there are codes with constant relative distance and constant dimension which have a basis tester, and in Proposition 14 we show in all codes have a basis tester whose query complexity equals to the dimension of the code plus one.

**Proposition 13** (The repetition code has a  $(2, 1)$ -strong uniform basis tester). *For any finite field  $\mathbb{F}$  and constant  $c \in \mathbb{N}^+$  there exists a  $[n = cm, k = c, d = m]_{\mathbb{F}}$ -code  $\mathcal{C}$  which has a  $(2, 1)$ -strong basis tester.*

*Proof:* Let  $\mathcal{C}$  be the  $[n = cm, k = c, d = m]_{\mathbb{F}}$  repetition code where a  $c$ -symbol message  $a_1, \dots, a_c$  is encoded by repeating each symbol  $m$  times, i.e.,  $a_1, \dots, a_c \mapsto a_1^m, \dots, a_c^m$ . Consider the tester that compares a random position in a block to the first bit in the block. Formally, the tester is defined by the uniform distribution over the following set  $B$  of words of weight 2 :  $B = \{e_{im+1} - e_{im+j} \mid i \in \{0, \dots, c-1\}, j \in \{2, \dots, m\}\}$ , where  $e_\ell$  has a 1 in the  $\ell$ th coordinate and is zero elsewhere.

It can be readily verified that  $B$  is a basis for  $\mathcal{C}^\perp$ , has query complexity 2 and rejects a word  $w$  with probability  $\delta(w, \mathcal{C})$  because if the rejection probability is  $\epsilon$  this means that at most an  $\epsilon$  fraction of symbols need to be changed to reach a word that is constant on each of its  $c$  blocks.  $\blacksquare$

**Proposition 14** (Every code has a basis tester with large query complexity). *Let  $\mathbb{F}$  be a finite field and  $\mathcal{C}$  be a  $[n, k, d]_{\mathbb{F}}$  code. Then  $\mathcal{C}$  has a  $(k+1, 1)$  strong uniform basis tester.*

*Proof:* Assume without loss of generality the first  $k$  entries of a codeword are message bits. This means that after querying the first  $k$  symbols of a word  $w_1, \dots, w_k$ , one can interpolate to obtain any other symbol of the codeword that is the encoding of the message  $(w_1, \dots, w_k)$ . For  $k < i \leq n$  let  $u_i$  be the constraint that queries the first  $k$  bits of  $w$  and accepts iff  $w_i$  is equal to the  $i$ th symbol of the encoding of  $(w_1, \dots, w_k)$ . It can be readily that  $B = \{u_{k+1}, \dots, u_n\}$  is a basis for  $\mathcal{C}^\perp$  and has query complexity  $k+1$ .

Consider the soundness of the uniform tester over  $B$ . If  $\Pr[T[w] = \text{reject}] \leq \rho$  then  $w$  is  $\rho$ -close to the codeword of  $\mathcal{C}$  that is the encoding of  $(w_1, \dots, w_k)$ , implying that  $\delta(w, \mathcal{C}) \leq \rho$ . ■

## VII. EVERY BINARY CODE CAN BE TESTED WITH REDUNDANCY LESS THAN $2 \dim(\mathcal{C}^\perp)$

We show that every binary linear code  $\mathcal{C}$  can be tested with linear redundancy, by proving the following statement. We point out that [5] implicitly showed already that every code can be tested with a linear amount of redundancy. The added value of the following statement is that it shows that the amount of redundancy can be as small as twice the dimension of  $\mathcal{C}^\perp$ .

**Proposition 15** ( $2 \dim(\mathcal{C}^\perp)$  redundancy is sufficient for testing any LTC). *If  $\mathcal{C}$  is a  $[n, k, d]_{\mathbb{F}_2}$  code that is a  $(q, \varepsilon, \rho)$ -LTC and  $\varepsilon \leq \delta(\mathcal{C})/3$ , then  $\mathcal{C}$  has a  $(\frac{10q}{\rho}, \varepsilon, \frac{1}{100})$ -tester whose support is over a set  $U$  of size at most  $3 \dim(\mathcal{C}^\perp)$  and additionally  $\dim(U) \geq \dim(\mathcal{C}^\perp) - 3\varepsilon n$ .*

**Remark 16.** Inspection of the proof of Proposition 15 reveals that  $\mathcal{C}$  can be tested by a  $(c \cdot q, \varepsilon, 1/c)$ -tester whose support is over  $U$  of size  $\leq (4 \ln 2 + \eta) \cdot (n - k)$  for any  $\eta > 0$ , where  $c > 1$  is a constant that depends on  $\eta$  and goes to infinity as  $\eta$  goes to 0. Recalling  $4 \ln 2 = 2.77258\dots$ , we preferred to round this constant up to the closest integer in the statement of the proposition above.

We point out that the support of a non-strong tester need not span  $\mathcal{C}^\perp$ . However, the lower bound on  $\dim(U)$  stated above implies that every tester's support must at least span a large subspace of  $\mathcal{C}^\perp$ . The proof of this proposition follows immediately from the following two claims.

**Claim 17.** *If  $\mathcal{C} \subseteq \mathbb{F}_2^n$  is a  $[n, k, d]_{\mathbb{F}_2}$ -code that is a  $(q, \varepsilon, \rho)$ -LTC, then it has a  $(\frac{10q}{\rho}, \varepsilon, \frac{1}{100})$ -tester whose support is over a set  $U$  of size at most  $3 \dim(\mathcal{C}^\perp)$ .*

**Claim 18.** *Let  $T$  be a  $(q, \varepsilon, \rho)$ -tester for a linear code  $\mathcal{C} \subseteq \mathbb{F}_2^n$  such that  $\varepsilon \leq \frac{\delta(\mathcal{C})}{3}$ . Let  $U \subseteq \mathcal{C}_{\leq q}^\perp$  denote the support of  $T$ . Then  $\dim(U) \geq \dim(\mathcal{C}^\perp) - 3\varepsilon n$ .*

In the remainder of this section we prove these two claim. Let us state a couple of inequalities in probability that will be used later on in the proof.

**Claim 19** (Chernoff Bound). *If  $X = \sum_{i=1}^m X_i$  is a sum of independent  $\{0, 1\}$ -valued random variables, where  $\Pr[X_i = 1] = \gamma$ , then*

$$\Pr[X < (1 - \sigma)\gamma m] \leq \exp(-\sigma^2 \gamma m / 2).$$

**Claim 20.** *If  $X = \sum_{i=1}^m X_i$  is a sum of independent  $\{0, 1\}$ -valued random variables, where  $\Pr[X_i = 1] = \gamma$ , then*

$$\Pr[X \equiv 0 \pmod{2}] \leq \frac{1}{2}(1 + \exp(-2\gamma m)).$$

*Proof:* Let  $Y_i = (-1)^{X_i}$  and let  $Y = \prod_{i=1}^m Y_i$ . Notice  $X \equiv 0 \pmod{2}$  iff  $Y = 1$ . Since  $Y$  is the product of independent random variables we have

$$\begin{aligned} \Pr[X \equiv 0 \pmod{2}] &= \mathbf{E}\left[\frac{1}{2}(1 + Y)\right] \\ &= \frac{1}{2}(1 + \prod_{i=1}^m \mathbf{E}[Y_i]) = \frac{1}{2}(1 + (1 - 2\gamma)^m) \\ &\leq \frac{1}{2}(1 + e^{-2\gamma m}). \end{aligned}$$

*Proof of Claim 17:* Let  $t = \frac{10}{\rho}$  and  $m = 3 \dim(\mathcal{C}^\perp) = 3(n - k)$ . Let  $T$  be the assumed  $(q, \varepsilon, \rho)$  tester for  $\mathcal{C}$ . Pick  $U = \{u_1, \dots, u_m\}$  where each  $u_i$  is obtained by taking the sum of  $t$  independent samples from  $T$ .  $U$  is a multiset and the distribution  $p$  associated with our tester is the uniform distribution over  $U$ . The query complexity of  $U$  is bounded by  $tq = \frac{10q}{\rho}$ .

To analyze soundness, fix a word  $w$  that is  $\varepsilon$ -far from  $\mathcal{C}$ . Let  $X_i$  be the indicator random variable for the event  $\langle w, u_i \rangle \neq 0$ . By Claim 20 it holds that

$$\Pr[X_i = 0] \leq \frac{1}{2}(1 + e^{-2\rho t})$$

and

$$\Pr[X_i = 1] \geq \frac{1}{2}(1 - e^{-2\rho t})$$

Let  $U_{\text{bad}} = \{u \in U \mid \langle u, w \rangle \neq 0\}$ . Then by the Chernoff bound (Claim 19) we have

$$\Pr\left[\frac{|U_{\text{bad}}|}{m} < \frac{1}{100}\right] \leq e^{-0.98^2(\frac{1}{2}(1 - e^{-2\rho t}))m/2}$$

We take a union bound over all words that are  $\varepsilon$ -far from  $\mathcal{C}$ . Notice that  $\mathbb{F}_2^n$  can be partitioned into  $2^{n-k}$  affine shifts of (the linear space)  $\mathcal{C}$ . For each such affine shift, which has the form  $v + \mathcal{C} = \{v + c \mid c \in \mathcal{C}\}$ , the probability of rejecting any two words from  $v + \mathcal{C}$  is equal, because they differ only by a word from  $\mathcal{C}$  which has inner product 0 with all tests. Thus, it suffices to take a union bound over one representative per affine shift, and there are at most  $2^{n-k}$  of them.

Continuing with the proof, the probability that there exists a  $\varepsilon$ -far word that is rejected with probability less than  $\frac{1}{100}$  is at most

$$e^{-0.98^2(\frac{1}{2}(1 - e^{-2\rho t}))m/2} \cdot 2^{n-k}$$



$$= e^{-0.98^2(\frac{1}{2}(1-e^{-2\rho t}))m/2+\ln(2)(n-k)}$$

We have

$$e^{-0.98^2(\frac{1}{2}(1-e^{-2\rho t}))m/2+\ln(2)(n-k)} < 1 \quad \text{if} \\ -0.98^2(\frac{1}{2}(1-e^{-2\rho t}))m/2 + \ln(2)(n-k) < 0.$$

By construction we have  $m > 2.95(n-k)$ . So

$$m > \frac{2.773(n-k)}{0.98^3} \Rightarrow \\ m > \frac{2.773(n-k)}{0.98^2} \Rightarrow \\ ((1-e^{-2\rho t}))m > \frac{4\ln(2)(n-k)}{0.98^2} \Rightarrow \\ -0.98^2(\frac{1}{2}(1-e^{-2\rho t}))m/2 + \ln(2)(n-k) < 0$$

Hence we showed that there is a positive probability to pick the set  $U$  such that every  $\varepsilon$ -far word is rejected with probability at least  $\frac{1}{100}$  and this completes the proof. ■

*Proof of Claim 18:* Assume by way of contradiction that  $\dim(U) < \dim(\mathcal{C}^\perp) - 3\varepsilon n$ . We call a word  $w$  a *coset leader* if  $w$  has minimal weight in  $w + \mathcal{C} = \{w + c \mid c \in \mathcal{C}\}$ . (If there is more than one minimal weight word in  $w + \mathcal{C}$  pick arbitrarily one of them to be the coset leader.) The proof of Proposition 9 implies that if  $w$  is a coset leader then  $\frac{\text{wt}(w)}{n} = \delta(w, \mathcal{C})$ .

Let

$$V = \{w \in F_2^n \setminus \mathcal{C} \mid \forall u \in U : \langle u, w \rangle = 0\}$$

and  $w$  is a coset leader of  $w + \mathcal{C}$ ,

i.e.  $V$  contains all non-codewords that are accepted by all tests in  $U$ . We have  $\dim(V) \geq 3\varepsilon n$  and thus  $|\bigcup_{v \in V} (\text{supp}(v))| \geq 3\varepsilon n$ . In addition for all  $v \in V$  we have  $\text{supp}(v) < \varepsilon n$  because

$$\Pr[T[v] = \text{reject}] = 0.$$

Let  $w_1, \dots, w_s$  be an arbitrary ordering of the elements of  $V$ . Let  $\mu(\ell)$  the maximal size of an element in  $\text{span}(w_1, \dots, w_\ell)$ . We have  $\mu(1) \leq \varepsilon n$  and  $\mu(s) \geq \frac{3}{2}\varepsilon n$  because the expected size of a word in  $\text{span}(V)$  is (exactly)  $\frac{1}{2}|\bigcup_{w \in V} (\text{supp}(w))|$ . Finally, we have  $\mu(\ell+1) < \mu(\ell) + \varepsilon n$ . We conclude there must exist  $\ell$  for which  $\varepsilon n < \mu(\ell) \leq 2\varepsilon n$ . Let  $w'$  be a word of maximal size in  $\text{span}(w_1, \dots, w_\ell)$ . We see that  $w'$  is  $\varepsilon$ -far from  $\mathcal{C}$  but accepted by  $T$  with probability 1, a contradiction. ■

## ACKNOWLEDGMENTS

We would like to thank Oded Goldreich for many valuable discussions including raising the question as to whether one can show LTCs can perform as well as random codes (a question that partly inspired this work, though not resolved yet). We also thank Oded and the anonymous referees for valuable comments on an earlier version of this article. We would like to thank Or Meir for pointers to the literature.

## REFERENCES

- [1] N. Alon, J. Bruck, J. Naor, M. Naor, and R. M. Roth, "Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs," *IEEE Transactions on Information Theory*, vol. 38, no. 2, p. 509, 1992.
- [2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, "Proof verification and the hardness of approximation problems," *Journal of the ACM*, vol. 45, no. 3, pp. 501–555, May 1998.
- [3] S. Arora and S. Safra, "Probabilistic checking of proofs: A new characterization of NP," *Journal of the ACM*, vol. 45, no. 1, pp. 70–122, Jan. 1998.
- [4] L. Babai, A. Shpilka, and D. Stefankovic, "Locally testable cyclic codes," in *Proceedings: 44th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2003, 11–14 October 2003, Cambridge, Massachusetts*, IEEE, Ed. IEEE Computer Society Press, 2003, pp. 116–125.
- [5] M. Bellare, O. Goldreich, and M. Sudan, "Free bits, PCPs, and nonapproximability—towards tight results," *SIAM Journal on Computing*, vol. 27, no. 3, pp. 804–915, Jun. 1998.
- [6] Y. Ben-Haim and S. Litsyn, "Upper bounds on the rate of LDPC codes as a function of minimum distance," *IEEE Transactions on Information Theory*, vol. 52, no. 5, pp. 2092–2100, 2006. [Online]. Available: <http://doi.ieeecomputersociety.org/10.1109/TIT.2006.872972>
- [7] E. Ben-Sasson, O. Goldreich, P. Harsha, M. Sudan, and S. P. Vadhan, "Robust PCPs of proximity, shorter PCPs, and applications to coding," *SIAM Journal on Computing*, vol. 36, no. 4, pp. 889–974, 2006.
- [8] E. Ben-Sasson, O. Goldreich, and M. Sudan, "Bounds on 2-query codeword testing," in *RANDOM-APPROX*, ser. Lecture Notes in Computer Science, vol. 2764. Springer, 2003, pp. 216–227. [Online]. Available: <http://springerlink.metapress.com/openurl.asp?genre=article&issn=0302-9743&volume=2764&page=216>

- [9] E. Ben-Sasson, P. Harsha, O. Lachish, and A. Matsliah, "Sound 3-query PCPPs are long," in *ICALP (I)*, ser. Lecture Notes in Computer Science, vol. 5125. Springer, 2008, pp. 686–697. [Online]. Available: [http://dx.doi.org/10.1007/978-3-540-70575-8\\_56](http://dx.doi.org/10.1007/978-3-540-70575-8_56)
- [10] E. Ben-Sasson, P. Harsha, and S. Raskhodnikova, "Some 3CNF properties are hard to test," *SIAM Journal on Computing*, vol. 35, no. 1, pp. 1–21, 2005. [Online]. Available: [http://epubs.siam.org/SICOMP/volume-35/art\\_44544.html](http://epubs.siam.org/SICOMP/volume-35/art_44544.html)
- [11] E. Ben-Sasson and M. Sudan, "Simple PCPs with poly-log rate and query complexity," in *STOC*. ACM, 2005, pp. 266–275. [Online]. Available: <http://doi.acm.org/10.1145/1060590.1060631>
- [12] E. Ben-Sasson and M. Sudan, "Robust locally testable codes and products of codes," *Random Struct. Algorithms*, vol. 28, no. 4, pp. 387–402, 2006. [Online]. Available: <http://dx.doi.org/10.1002/rsa.20120>
- [13] I. Dinur, "The PCP theorem by gap amplification," *Journal of the ACM*, vol. 54, no. 3, pp. 12:1–12:44, Jun. 2007.
- [14] I. Dinur and O. Reingold, "Assignment testers: Towards a combinatorial proof of the PCP theorem," *SIAM Journal on Computing*, vol. 36, no. 4, pp. 975–1024, 2006. [Online]. Available: <http://dx.doi.org/10.1137/S0097539705446962>
- [15] I. Dinur, M. Sudan, and A. Wigderson, "Robust local testability of tensor products of LDPC codes," in *APPROX-RANDOM*, ser. Lecture Notes in Computer Science, vol. 4110. Springer, 2006, pp. 304–315. [Online]. Available: [http://dx.doi.org/10.1007/11830924\\_29](http://dx.doi.org/10.1007/11830924_29)
- [16] O. Goldreich, "Short locally testable codes and proofs (survey)," *Electronic Colloquium on Computational Complexity (ECCC)*, no. 014, 2005. [Online]. Available: <http://eccc.hpi-web.de/eccc-reports/2005/TR05-014/index.html>
- [17] O. Goldreich and M. Sudan, "Locally testable codes and PCPs of almost-linear length," *Journal of the ACM*, vol. 53, no. 4, pp. 558–655, Jul. 2006.
- [18] E. Grigorescu, T. Kaufman, and M. Sudan, "2-transitivity is insufficient for local testability," in *IEEE Conference on Computational Complexity*. IEEE Computer Society, 2008, pp. 259–267. [Online]. Available: <http://doi.ieeeecomputersociety.org/10.1109/CCC.2008.31>
- [19] V. Guruswami, "On 2-query codeword testing with near-perfect completeness," in *ISAAC*, ser. Lecture Notes in Computer Science, vol. 4288. Springer, 2006, pp. 267–276. [Online]. Available: [http://dx.doi.org/10.1007/11940128\\_28](http://dx.doi.org/10.1007/11940128_28)
- [20] T. Kaufman and M. Sudan, "Sparse random linear codes are locally decodable and testable," in *FOCS*. IEEE Computer Society, 2007, pp. 590–600. [Online]. Available: <http://doi.ieeeecomputersociety.org/10.1109/FOCS.2007.65>
- [21] T. Kaufman and M. Sudan, "Algebraic property testing: the role of invariance," in *STOC*. ACM, 2008, pp. 403–412. [Online]. Available: <http://doi.acm.org/10.1145/1374376.1374434>
- [22] O. Meir, "Combinatorial construction of locally testable codes," in *STOC*. ACM, 2008, pp. 285–294. [Online]. Available: <http://doi.acm.org/10.1145/1374376.1374419>
- [23] R. M. Roth, *Introduction to coding theory*. pub-CAMBRIDGE:adr: Cambridge University Press, 2006.
- [24] D. A. Spielman, "Linear-time encodable and decodable error-correcting codes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1723–1731, 1996.
- [25] M. Sudan, "Algorithmic introduction to coding theory, lecture notes," 2001. [Online]. Available: <http://theory.csail.mit.edu/~madhu/FT01/>
- [26] P. Valiant, "The tensor product of two codes is not necessarily robustly testable," in *APPROX-RANDOM*, ser. Lecture Notes in Computer Science, vol. 3624. Springer, 2005, pp. 472–481. [Online]. Available: [http://dx.doi.org/10.1007/11538462\\_40](http://dx.doi.org/10.1007/11538462_40)