Electronic Colloquium on Computational Complexity, Report No. 98 (2007)

Breaking the ϵ-Soundness Bound of the Linearity Test over GF(2)

Tali Kaufman* Simon Litsyn† Ning Xie‡

Abstract

For Boolean functions that are ϵ-far from the set of linear functions, we study the lower bound on the rejection probability (denoted $\text{REJ}(\epsilon)$) of the linearity test suggested by Blum, Luby and Rubinfeld. The interest in this problem is partly due to its relation to PCP constructions and hardness of approximating some NP-hard problems. It seems that the problem of lower bounding $\text{REJ}(\epsilon)$ becomes more difficult as ϵ approaches 1/2.

The previously best bounds for $\text{REJ}(\epsilon)$ were obtained by Bellare, Coppersmith, Håstad, Kiwi and Sudan. They used Fourier analysis to show that $\text{REJ}(\epsilon) \geq \epsilon$ for every $0 \leq \epsilon \leq \frac{1}{2}$. They also conjectured that this bound might not be tight for ϵ's which are close to 1/2. In this paper we show that this indeed is the case. Specifically, we improve the lower bound of $\text{REJ}(\epsilon) \geq \epsilon$ by an additive constant that depends only on ϵ: $\text{REJ}(\epsilon) \geq \epsilon + \min\{1376\epsilon^3(1-2\epsilon)^{12}, \frac{1}{4}\epsilon(1-2\epsilon)^4\}$, for every $0 \leq \epsilon \leq \frac{1}{2}$. Our analysis is based on a relationship between $\text{REJ}(\epsilon)$ and the weight distribution of a coset of the Hadamard code. We use both Fourier analysis and coding theory tools to estimate this weight distribution.

*CSAIL, MIT, Cambridge, MA 02139. E-mail: kaufmant@mit.edu.

†Department of Electrical Engineering-Systems, Tel Aviv University, Tel Aviv 69978, ISRAEL. E-mail: litsyn@eng.tau.ac.il.

‡CSAIL, MIT, Cambridge, MA 02139. E-mail: ningxie@gmail.com. Research done while the author was at State Univ. of New York at Buffalo and visiting CSAIL, MIT. Partially supported by NSF grant 0514771.

ISSN 1433-8092

1 Introduction

A function $f : \{0,1\}^m \to \{0,1\}$ is called *linear* if for all $x, y \in \{0,1\}^m$, $f(x) + f(y) = f(x+y)$. A function f is said to be ϵ-far from linear functions if one needs to change f's value on at least an ϵ-fraction of its domain to make f linear. Blum, Luby and Rubinfeld [5] considered the following randomized algorithm (henceforth referred to as "BLR test") to test if a function is linear : Given a function $f : \{0,1\}^m \to \{0,1\}$, choose uniformly at random $x, y \in \{0,1\}^m$ and reject if $f(x) + f(y) \neq f(x+y)$. It is obvious that the completeness of BLR test is one, i.e., if f is linear, then BLR test always accepts. However, the soundness analysis of the BLR test turned out to be highly complex. Indeed, various papers studied the following question: For all Boolean functions that are ϵ-far from linear functions, what is the minimum rejection probability of the BLR linearity tests. We denote this lower bound by $\text{REJ}(\epsilon)$. $\text{REJ}(\epsilon)$ is important not only because it is a natural combinatorial problem but also due to the fact that it is related to the hardness of approximating some NP-hard problems. Prior to the paper of Håstad on optimal in-approximation results [7], any improvement of $\text{REJ}(\epsilon)$ would yield improvements of some inapproximability results. See [1] for relevant discussions. After Håstad's celebrated results, there are less immediate implications of improvements for $\text{REJ}(\epsilon)$. However, since most recent PCP constructions are based on devising and analyzing some Long Code tests, and linearity test is still relevant to the analysis of the Long Code, it seems understanding $\text{REJ}(\epsilon)$ is an important question of its own right. The most interesting (and also most difficult) cases are those where $\frac{1}{4} \leq \epsilon < \frac{1}{2}$.

1.1 Related research

Blum, Luby and Rubinfeld [5] first suggested the BLR linearity test and showed that for every ϵ, $\text{REJ}(\epsilon) \geq \frac{2}{9}\epsilon$ based on a self-correction approach. Using a combinatorial argument, Bellare et al. [3] proved that $\text{REJ}(\epsilon) \geq 3\epsilon - 6\epsilon^2$. This bound is optimal for small ϵ but is very weak for ϵ's that are close to $\frac{1}{2}$. Bellare and Sudan [4] further showed that $\text{REJ}(\epsilon) \geq \frac{2}{3}\epsilon$ when $\epsilon \leq \frac{1}{3}$ and $\text{REJ}(\epsilon) \geq \frac{2}{9}$ when $\epsilon > \frac{1}{3}$. This series of works culminated in [1], where Fourier transform techniques found their first use in PCP-related analysis. The linear lower bound $\text{REJ}(\epsilon) \geq \epsilon$ as well as the so-called "Knee" bound, $\text{REJ}(\epsilon) \geq \frac{45}{128}$ when $\epsilon \geq \frac{1}{4}$, are all due to [1]. They also showed that all the known bounds are tight for $\epsilon \leq \frac{5}{16}$. In [1], numerical simulation results suggested that the lower bound $\text{REJ}(\epsilon) \geq \epsilon$ for $\epsilon > \frac{5}{16}$ may be improved, but not by too much. Kiwi [9] and Kaufman and Litsyn [8] gave alternative proofs for the fact that $\text{REJ}(\epsilon) \geq \epsilon$ for every ϵ (up to an additive term of $O(\frac{1}{n})$). Their proofs are more coding theory oriented. Specifically, the proofs are based on studying the weight distribution of the Hadamard code and its ϵ-away coset as well as various properties of Krawtchouk polynomials.

Apart from the lower bounds for the rejection probability, some other limits of the BLR test were also explored. Specifically, Trevisan [15] and Samorodnitsky and Trevisan [12] have studied so-called "amortized query complexity" of the BLR test. Let f be a function that is very far from linear functions. Amortized query complexity quantifies the trade-off between the number of points at which one needs to query f and the probability of accepting f as a linear function. Linearity tests with low amortized query complexity are useful in constructing PCP systems with low amortized query complexity.

Some research (see e.g. [13]) observed the similarity between the methodologies of BLR test analysis [1] and Roth's proof [11] that any subset of integers with positive density contains arith-

metic progressions of length 3 (see e.g. [11, 14, 6] for discussions about arithmetic progressions in dense sets). We hope that our new results on the limits of the linearity test may be of interest in this respect as well.

On the limits of Fourier for bounding binary functions When combined with previous bound on $\textsc{rej}(\epsilon)$ for $\epsilon \in (0, \frac{1}{4}]$, our result implies that the lower bound on $\textsc{rej}(\epsilon)$ obtained via Fourier analysis is sub-optimal for all $\epsilon \in (0, \frac{1}{2})$. We improve the lower bound by combining coding theory and Fourier analysis techniques. We believe that the suggested method is of independent interest in coding theory and may find other applications in enhancement of results on Boolean functions based on Fourier analysis.

1.2 The main result

In the following, we present our main result showing an improved bound for $\textsc{rej}(\epsilon)$. Specifically, we prove

Theorem 1. *Let* $\Delta(\gamma) = \frac{5\gamma}{8} - \frac{\gamma^2}{32}$. *For all* ϵ, $1/4 \leq \epsilon \leq 1/2$ *and for all* γ, $0 < \gamma \leq 1$,

$$\textsc{rej}(\epsilon) \geq \epsilon + \min\{4096(1-\Delta(\gamma))^3\epsilon^3(1-2\epsilon)^{12}, \frac{\gamma}{2}\epsilon(1-2\epsilon)^4\}.$$

As a simple corollary by plugging in $\gamma = 1/2$ and combine our new result with known bounds for $0 \leq \epsilon \leq \frac{1}{4}$ (i.e., $\textsc{rej}(\epsilon) \geq 3\epsilon - 6\epsilon^2$), we get

Corollary 2. *For all* ϵ, $0 \leq \epsilon \leq 1/2$,

$$\textsc{rej}(\epsilon) \geq \epsilon + \min\{1376\epsilon^3(1-2\epsilon)^{12}, \frac{1}{4}\epsilon(1-2\epsilon)^4\}.$$

Note that Theorem 1 improves upon $\textsc{rej}(\epsilon) \geq \epsilon$ by an additive *constant* that depends only on ϵ for all $\frac{1}{4} \leq \epsilon < \frac{1}{2}$. Our result improves over all previously known bounds for every ϵ, $\frac{45}{128} \leq \epsilon < \frac{1}{2}$, but only by a very small quantity. For example, for $\epsilon = 0.4$, our improvement is about 1.024×10^{-7}. We believe our bound can be further improved systematically (note that our current approach already gives bounds better than that stated in the Main Theorem for ϵ's such that $1/(1-2\epsilon)^2$ is close to integers). However, as the numerical results shown in [1], one can not expect to see too much improvement over $\textsc{rej}(\epsilon) \geq \epsilon$. Note that our improvement over $\textsc{rej}(\epsilon) \geq \epsilon$ vanishes at $\epsilon = \frac{1}{2}$. This is indeed as expected since we know that $\textsc{rej}(\frac{1}{2}) = \frac{1}{2}$.

1.3 Proof overview

The proof has three key ingredients. We will use C to denote the Hadamard code of length n whose codewords are exactly the set of all linear functions.

The coset code $C+v$ There are two equivalent ways of viewing the BLR test: one is to think f as a Boolean function mapping $\{0,1\}^m$ to $\{0,1\}$ and the BLR test simply picks x and y uniformly at random and check if $f(x) + f(y) = f(x+y)$. This functional viewpoint leads naturally to the beautiful Fourier analysis approach of [1], which shows that $\textsc{rej}(\epsilon)$ can be exactly expressed as a

cubic sum of Fourier coefficients of the function $(-1)^f$. Another way to study the BLR test, first suggested in [9] and followed by [8], treats f as a codeword v of length n with $n = 2^m$. (Due to this fact, from now on, we will use codeword v and function f interchangeably.) Since the set of linear functions may be viewed as the set of codewords of the Hadamard code C. BLR test can be viewed as picking a random weight-3 codeword from $C^\perp$ and check if it is orthogonal to v. We combine these two viewpoints together by reinterpreting the Fourier analytic result in the coding theoretic setting. Our simple but important observation is that the Fourier coefficients of f are equivalent to the weights of the codewords in a coset of C. Therefore $\text{REJ}(\epsilon)$ can be expressed as a simple function of the weight distribution of the code $C + v$, where $C + v$ is an ϵ-away coset of the Hadamard code C. To make this clear, we remind the reader that the weight distribution of a code C is a set of integers that represent the numbers of codewords in C of different weights, where the weight of a codeword is the number of coordinates at which the codeword is non-zero. A vector v is ϵ-far from a code C if one needs to change at least an ϵ-fraction of v's bits to make it belong to C. An ϵ-away coset of C is obtained by adding a vector v that is ϵ-far from C to every codeword in C.

Maximization Problem In order to obtain a lower bound on a function that involves the weight distribution of $C + v$, we reformulate our problem as a Maximal Sum of Cubes Problem, in which we look for an upper bound on the sum of cubes of a set of integers under certain constraints. The bound $\text{REJ}(\epsilon) = \epsilon$ corresponds to a simple optimal configuration in which all the codewords of $C + v$ sit at weight $\frac{1}{2}n$ except a constant number $(\frac{1}{(1-2\epsilon)^2})$ of them sit at weight ϵn. Moreover, this is the unique configuration that meets the bound $\text{REJ}(\epsilon) = \epsilon$. Any deviation from the optimal configuration implies an improved lower bound on $\text{REJ}(\epsilon)$. Our strategy thus is to show that this optimal weight distribution is not achievable due to some special properties of the code $C + v$. In particular, we will focus on the following two ways in which the optimal configuration may break down:

1. There exists a *heavy codeword* in $C + v$, i.e. a codeword of weight larger than $\frac{n}{2}$.
2. The number of codewords in $C + v$ of weight at most $(\epsilon + \eta)n$ is less than $\frac{1}{(1-2\epsilon)^2}$, where η is a positive number.

A natural tool to show that one of the above properties holds is the Johnson Bound. Roughly speaking, the well-known Johnson bound offers a bound on the maximum number of codewords of a specific weight in a code with some specific minimum distance. However, it turns out that Johnson bound is met *exactly* by the code $C+v$ at weight ϵn, and we fail to get any improvement by applying it directly to $C + v$. The way we overcome this is by considering a new code $C|_{\mathcal{V}}$ and applying a slightly stronger variant of the commonly used Johnson bound which enables us to bound the number of codewords of *at least* (or *at most*) a specific weight. This turns out to be crucial in our analysis.

From the code $C + v$ to the code $C|_{\mathcal{V}}$ We consider the code $C|_{\mathcal{V}}$ of block length $n' = \epsilon n$, obtained from C by restricting it to the ϵn non-zero coordinates of v. This code is a linear code. It has the same number of codewords as the original code $C + v$. More precisely, we show that if it contains fewer codewords then an improved lower bound on $\text{REJ}(\epsilon)$ is immediate. A

nice property of this new code is that there is a one-to-one correspondence between the weight of a codeword in $C|_{\mathcal{V}}$ and the weight of the corresponding codeword in $C + v$. Since $C|_{\mathcal{V}}$ is a linear code, its minimum distance equals the minimum weight of its codewords. If this minimum weight is small, then by the one-to-one relation between the weights of $C + v$ and that of $C|_{\mathcal{V}}$, the heaviest codeword in $C + v$ will have a large weight, which yields an improved lower bound for $\text{REJ}(\epsilon)$ according to Condition 1 from above. However, if the maximum weight of $C + v$ is small, or equivalently, the minimum distance of $C|_{\mathcal{V}}$ is large, then by applying the Johnson bound to $C|_{\mathcal{V}}$, we get that the number of codewords lying between weight ϵn and $(\epsilon + \eta)n$ in $C + v$ is less than the optimal bound $(\frac{1}{(1-2\epsilon)^2})$, which also yields an improved lower bound for $\text{REJ}(\epsilon)$ by Condition 2 mentioned before.

The intuitive reason that we gain from applying the Johnson bound to $C|_{\mathcal{V}}$ rather than to $C + v$, is because the block length of $C|_{\mathcal{V}}$ is much smaller than the block length of $C + v$, but the number of codewords in $C|_{\mathcal{V}}$ is the same as $C + v$[1].

The relations between the three codes in consideration, namely C, $C + v$, and $C|_{\mathcal{V}}$ (for a code C and a vector v that is ϵ-far from C), as well as the idea of looking at a restricted code of smaller block length in order to get better coding bounds, might have other applications.

1.4 Outline

In section 2 we show $\text{REJ}(\epsilon)$ can be expressed as a function of the weight distribution of a coset of the Hadamard code. Then we reformalize the problem of lower bounding $\text{REJ}(\epsilon)$ as a maximization problem in section 3. In section 4 we study the weight distribution of a restricted code of the coset code and then sketch the proof of the Main Theorem in section 5. Several technical claims appear in the Appendix.

2 The coset code $C + v$

First we introduce some notation. Let v be a vector in $\{0,1\}^n$. The weight of v, denoted $\text{wt}(v)$ is the number of non-zero bits in v. A code C of block length n is a subset of $\{0,1\}^n$. C is called a linear code if C is a linear subspace. Let $u, v \in \{0,1\}^n$. The *distance* between u and v is defined to be the number of bits at which they disagree: $\text{dist}(u,v) = |\{i \in [n] | u(i) \neq v(i)\}| = \text{wt}(u - v)$. The minimum distance of a code C is $\min_{u,v \in C} \text{dist}(u,v)$. If C is a linear code, then the minimum distance of C equals the minimum weight of codewords in C. Let C be a code of block length n. The distance of $v \in \{0,1\}^n$ from code C is the minimum distances between v and codewords in C, i.e., $\text{dist}(v, C) \stackrel{\text{def}}{=} \min_{c \in C} \text{dist}(v, c)$.

Let C be a linear code of block length n and let $v \in \{0,1\}^n$ such that $v \notin C$, the *v-coset* of C is $C + v \stackrel{\text{def}}{=} \{c + v | c \in C\}$. Note that $|C + v| = |C|$. The *weight distribution* or *spectrum* of C is

[1]The reason we are able to improve the bound $\text{REJ}(\epsilon) \geq \epsilon$ by a constant is more subtle: For $\frac{1}{4} \leq \epsilon \leq \frac{1}{2}$, there is a "reciprocal" relationship between the *relative weights* of codeword in C and corresponding codeword in $C|_{\mathcal{V}}$; that is, the smaller the relative weight in C, the larger the relative weight in $C|_{\mathcal{V}}$, and vice versa. Note that the denominator of the expression in Johnson bound is $\frac{d}{n} - 2\frac{w}{n}(1 - \frac{w}{n})$ after dividing by n^2. Therefore Johnson bound will give better bounds when $\frac{w}{n}(1 - \frac{w}{n})$ gets smaller, or, when w/n is very close to either 0 or 1. By switching from C to $C|_{\mathcal{V}}$, $\frac{w}{n}$ is mapped to $\frac{w'}{n'}$. The advantage of changing to $C|_{\mathcal{V}}$ is that it makes the distance between $\frac{w'}{n'}$ and 1 smaller than the distance between $\frac{w}{n}$ and zero. This advantage disappears at $\epsilon = 1/2$, therefore we get no improvement at that point, as expected.

$B^C = (B_0^C, B_1^C, \cdots, B_n^C)$, where $B_i^C = |\{c \in C | \mathrm{wt}(c) = i\}|$.

With an abuse of notation, in the following, we will use C to denote the Hadamard code and $C^\perp$ to denote its dual Hamming code.

Recall that a function $\ell : \{0,1\}^m \to \{0,1\}$ is linear if for all $x, y \in \{0,1\}^m$, $\ell(x)+\ell(y) = \ell(x+y)$. An equivalent characterization is: ℓ is linear if and only if $\ell(x) = \alpha \cdot x = \sum_i^m \alpha_i x_i$ for some $\alpha \in \{0,1\}^m$, and we denote such a linear function by ℓ_α and denote the set of all such functions by LIN. Let $f, g : \{0,1\}^m \to \{0,1\}$. The (relative) distance between f and g is defined to be the fraction of points at which they disagree: $\mathrm{dist}(f,g) \stackrel{\mathrm{def}}{=} \Pr_{x\in\{0,1\}^m}[f(x) \neq g(x)]$. The distance between a function f and linear functions is the minimum distance between f and any linear function: $\mathrm{dist}(f, \mathrm{LIN}) \stackrel{\mathrm{def}}{=} \min_{g\in \mathrm{LIN}} \mathrm{dist}(f,g)$. A function f is said to be ϵ-far from linear functions if its distance from linear functions is at least ϵ.

Note that the set of 2^m functions, $\psi_\alpha(x) = (-1)^{\alpha\cdot x}$, forms an orthonormal basis for the vector space of the set of functions $f : \{0,1\}^m \to \mathbb{R}$. Therefore $f(x)$ can be expanded as

$$f(x) = \sum_\alpha \hat{f}_\alpha \psi_\alpha(x),$$

where $\hat{f}_\alpha = \langle f, \psi_\alpha \rangle \stackrel{\mathrm{def}}{=} \frac{1}{2^m} \sum_{x\in\{0,1\}^m} f(x)\psi_\alpha(x)$ is called the α-th Fourier coefficient of f. Define $h(x) = (-1)^{f(x)}$. Using Fourier analytic tools, Bellare et al. proved that, if $\mathrm{dist}(f, \mathrm{LIN}) = \epsilon$ then

Lemma 3 ([1])**.** $\textsc{rej}(\epsilon) = \frac{1}{2}\left(1 - \sum_{\alpha\in\{0,1\}^m} \hat{h}_\alpha^3\right)$.

One can encode f as an $n = 2^m$ bit codeword $v \in \{0,1\}^n$ by enumerating all its values on the Boolean cube. The same encoding applied to the set of linear functions $\{\ell_\alpha\}$ gives rise to the Hadamard code C, in which we denote the codeword corresponding to ℓ_α by c_α. Then we have the following coding theoretic formula for $\textsc{rej}(\epsilon)$:

Lemma 4. $\textsc{rej}(\epsilon) = \frac{1}{2}\left(1 - \frac{1}{n^3}\sum_{i=0}^n B_i^{C+v}(n-2i)^3\right)$.

Proof. By definition of Fourier coefficient,

$$\hat{h}_\alpha = \langle h, \psi_\alpha \rangle = \langle (-1)^f, (-1)^{\ell_\alpha} \rangle = \langle (-1)^v, (-1)^{c_\alpha} \rangle = \frac{1}{2^m} \sum_{x\in\{0,1\}^m} (-1)^{v(x)+c_\alpha(x)}$$

$$= \Pr_x[v(x) = c_\alpha(x)] - \Pr_x[v(x) \neq c_\alpha(x)] = 1 - \frac{2\mathrm{dist}(v, c_\alpha)}{n} = \frac{n - 2\mathrm{wt}(v + c_\alpha)}{n},$$

where in the last step we use the fact that, for binary vectors u and v, $\mathrm{dist}(u,v) = \mathrm{wt}(u-v) = \mathrm{wt}(u+v)$. Lemma 3 now gives

$$\textsc{rej}(\epsilon) = \frac{1}{2}\left(1 - \sum_{\alpha\in\{0,1\}^m} \hat{h}_\alpha^3\right) = \frac{1}{2}\left(1 - \sum_{\alpha\in\{0,1\}^m} \frac{(n - 2\mathrm{wt}(v+c_\alpha))^3}{n^3}\right)$$

$$= \frac{1}{2}\left(1 - \sum_{c\in C} \frac{(n - 2\mathrm{wt}(v+c))^3}{n^3}\right)$$

$$= \frac{1}{2}\left(1 - \frac{\sum_{i=0}^n B_i^{C+v}(n-2i)^3}{n^3}\right),$$

where in the final step we change summation over codewords in C to summation over weights of the codewords in $C + v$. ■

This relation between the Fourier coefficients of $(-1)^f$ and the weight distribution of coset code $C + v$ seems to be new and may find applications in other places.

Since $\text{REJ}(\epsilon)$ is now expressed as a weight distribution of the coset code $C + v$, our next step is to study how the codewords in $C + v$ are distributed so that to make the rejection probability minimum.

3 Maximization problem

Note that we can rewrite Lemma 4 as

$$\text{REJ}(\epsilon) = \frac{1}{2} - \frac{\sum_{c_i \in C} (n - 2\text{wt}(v + c_i))^3}{2n^3} = \frac{1}{2} - \frac{1}{2n^3} \sum_{i=0}^{n-1} x_i^3,$$

where $x_i = n - 2\text{wt}(c_i + v)$, for $c_i \in C$, $0 \leq i \leq n-1$. Hence our goal of getting a better *lower* bound than ϵ for $\text{REJ}(\epsilon)$ is equivalent to getting a better *upper* bound than $1 - 2\epsilon$ for $\frac{1}{n^3} \sum_{i=0}^{n-1} x_i^3$. This observation motivates the following measure of improvement (gain) and to reformulate the problem of lower bounding $\text{REJ}(\epsilon)$ as a Maximal Sum of Cubes Problem.

Definition 5. *Let $x_i = n - 2wt(c_i + v)$, for $c_i \in C$, $0 \leq i \leq n-1$. Define*

$$\text{GAIN}(\epsilon) = \frac{1}{n^3} \left((1 - 2\epsilon)n^3 - \sum_{i=0}^{n-1} x_i^3 \right).$$

Consequently, $\text{REJ}(\epsilon) = \epsilon + \frac{1}{2}\text{GAIN}(\epsilon)$.

Since v is ϵ-far from C, it follows that $x_i \leq (1 - 2\epsilon)n$, for all $0 \leq i \leq n-1$. We further observe another constraint on the set of integers $\{x_0, x_1, \ldots, x_{n-1}\}$ is that their Euclidean norm is n^2.

Claim 6. $\sum_{i=0}^{n-1} x_i^2 = n^2$.

This claim follows directly from the Parseval's equality. An alternative proof, based on the norm-preserving property of the Hadamard matrix, was given in [8].

As we will show in the next lemma, if these two constraints are the only constraints on $\{x_0, x_1, \ldots, x_{n-1}\}$, then the bound $\text{REJ}(\epsilon) \geq \epsilon$ is essentially optimal. However, as we will see in the next section, since the x_i's are related to the weight distribution of $C + v$, the properties of the code $C + v$ impose more constraints on x_i's, thus making this optimal bound unattainable.

Lemma 7. *Consider the following Maximal Sum of Cubes Problem: For a set of n integers $x_0, x_1, \cdots, x_{n-1}$, find the maximum of $x_0^3 + x_1^3 + \cdots + x_{n-1}^3$ under the constraints:*

$$x_0^2 + x_1^2 + \cdots + x_{n-1}^2 = n^2$$

$$\forall i : x_i \leq \alpha n, \text{ for some } \alpha \leq 1$$

The maximum is achieved when a $\frac{1}{\alpha^2}$ fraction of the x_i's are assigned the maximum value αn, and the rest are assigned the value zero. The maximum thus obtained is αn^3.

The proof of this lemma is deferred to Appendix B. Note that in our setting $x_i = n - 2\mathrm{wt}(c_i + v)$ so $\alpha = 1 - 2\epsilon$ and hence $\sum_{i=0}^{n-1} x_i^3 \leq (1-2\epsilon)n^3$.

We will employ the following two lemmas on $\textsc{gain}(\epsilon)$ to obtain improvement.

Lemma 8. *If there exists an x_i such that $x_i = -\delta n$ for some $\delta > 0$, then $\textsc{gain}(\epsilon) \geq \min\{2\delta^3, 2\alpha^3\}$.*

Proof. We first consider the case that $\delta \leq \alpha$. Note that $\{x_0, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_{n-1}\}$ satisfies all the constraints in the Maximal Sum of Cubes Problem, so we have

$$\alpha n^3 \geq \sum_{k=0, k\neq i}^{n-1} x_k^3 + (-x_i)^3 = \sum_{k=0}^{n-1} x_k^3 + 2|x_i|^3 = \sum_{k=0}^{n-1} x_k^3 + 2(\delta n)^3.$$

The case $\delta > \alpha$ can be proved by observing that when $x_i \leq -\alpha n$, then $\sum_{i=0}^{n-1} x_i^3$ decreases as x_i decreases. ∎

Lemma 9. *Let $\eta > 0$. If the number of x_i's such that $x_i \geq (\alpha - \eta)n$ is at most $\lfloor \frac{1}{\alpha^2} \rfloor - 1$, then $\textsc{gain}(\epsilon) \geq \alpha^2 \eta$.*

Proof. Set $M = \lfloor \frac{1}{\alpha^2} \rfloor$. Let $\{y_1, \ldots, y_n\}$ be a permutation of $\{x_0, \ldots, x_{n-1}\}$ such that $\alpha n \geq y_1 \geq \cdots \geq y_n$. We have $y_1^2 + \cdots + y_n^2 = n^2$ and $y_M \leq (\alpha - \eta)n$. Define T to be: $T = y_1^2 + \cdots + y_{M-1}^2$. Then we have $T \leq (M-1)(\alpha n)^2 \leq (\frac{1}{\alpha^2} - 1)\alpha^2 n^2$, and $y_M^2 + \cdots + y_n^2 = n^2 - T$. Therefore,

$$\begin{aligned}\sum_{i=0}^{n-1} x_i^3 = \sum_{i=1}^{n} y_i^3 &\leq (\sum_{i=1}^{M-1} y_i^2)\alpha n + (\sum_{i=M}^{n} y_i^2)(\alpha - \eta)n = n^2(\alpha - \eta)n + \eta n T \\ &\leq n^2(\alpha - \eta)n + \eta n(\frac{1}{\alpha^2} - 1)\alpha^2 n^2 = \alpha n^3 - \alpha^2 \eta n^3.\end{aligned}$$

∎

4 From the code $C + v$ to the code $C|_{\mathcal{V}}$

We denote by $\mathcal{V}$ the set of coordinates at which v is non-zero, i.e., $\mathcal{V} = \{i \in [n] | v(i) = 1\}$. Note that $|\mathcal{V}| = \mathrm{wt}(v)$. In the following we consider a code $C|_{\mathcal{V}}$ which will enable us to get some insight into the weight distribution of the code $C + v$.

First observe that, since we are only interested in the weight distribution of $C + v$, without loss of generality, we may assume that $\mathrm{wt}(v) = \epsilon n$. To see this, suppose that $c_v \in C$ is the closest codeword to v. Since $\mathrm{dist}(v, C) = \epsilon n$, v can be written as $v = c_v + v_{\epsilon n}$, with $\mathrm{wt}(v_{\epsilon n}) = \epsilon n$. Since C is a linear code, $C + v = \{c + v | c \in C\} = \{c + c_v + v_{\epsilon n} | c \in C\} = \{c' + v_{\epsilon n} | c' \in C\} = C + v_{\epsilon n}$, where $c' \stackrel{\text{def}}{=} c + c_v$.

Definition 10. *Let C be a code of block length n and $v \in \{0,1\}^n$ be a vector of weight ϵn. We define the code $C|_{\mathcal{V}}$ of block length ϵn to be the code obtained by restricting code C to the non-zero coordinates of v. For convenience of notation, we will use $D = C|_{\mathcal{V}}$ from now on.*

The following lemma shows that there is a one-to-one correspondence between the weight of $c_i + v$ and the weight of the corresponding codeword in D.

Lemma 11. *For $0 \le i \le n-1$, let c_i be the ith codeword in the Hadamard code C and $d_i \in D$ be the restriction of c_i to coordinates in $\mathcal{V}$. Let $x_i = n - 2wt(c_i + v)$, then*

$$x_i = \begin{cases} (1-2\epsilon)n, & \text{if } i = 0, \\ 4wt(d_i) - 2\epsilon n, & \text{otherwise.} \end{cases}$$

Proof. For $i = 0$, $\mathrm{wt}(c_0 + v) = \mathrm{wt}(v) = \epsilon n$, hence $x_0 = (1-2\epsilon)n$. Since C is a Hadamard code, for all $i > 0$, $\mathrm{wt}(c_i) = n/2$, i.e., there are $n/2$ ones and $n/2$ zeros in each codeword. For each $c_i \in C$, since there are $\mathrm{wt}(d_i)$ ones in $\mathcal{V}$, there are $n/2 - \mathrm{wt}(d_i)$ ones in $[n] \setminus \mathcal{V}$; this also holds for $c_i + v$, since v does not flip the bits at these coordinates. Since $|v| = \epsilon n$, there are $\epsilon n - \mathrm{wt}(d_i)$ zeros in $\mathcal{V}$ for c_i, therefore there are $\epsilon n - \mathrm{wt}(d_i)$ ones in $\mathcal{V}$ for $c_i + v$. It follows that $\mathrm{wt}(c_i + v) = n/2 - \mathrm{wt}(d_i) + \epsilon n - \mathrm{wt}(d_i) = (1/2 + \epsilon)n - 2\mathrm{wt}(d_i)$ and $x_i = 4\mathrm{wt}(d_i) - 2\epsilon n$. ■

Lemma 12. *Either D is a linear code or $\textsc{gain}(\epsilon) \ge 2(1-2\epsilon)^3$.*

Proof. Since D is a restriction of linear code C, D is a linear code if and only if all the codewords d_i in D are distinct. If D is not a linear code, then there exist $i \ne j$ such that $d_i = d_j$. This implies that there is a $k \ne 0$ such that $d_k = \vec{0}$. By Lemma 11, $x_k = -2\epsilon n$. Since $2\epsilon \ge 1 - 2\epsilon$, by Lemma 8, $\textsc{gain}(\epsilon) \ge 2(1-2\epsilon)^3$. ■

Since $2(1-2\epsilon)^3$ is always larger than the gain we are going to prove, from now on, we will focus on the case that D is a linear code. Let $n' = \epsilon n$ be the block length of D, and d be the minimum distance of D. Note that D contains n codewords. The following simple bound is useful.

Theorem 13 (Plotkin bound [10])**.** *Let C be a binary code of block length n and minimum distance d. If $d \ge n/2$, then C has at most $2n$ codewords.*

Now we have

Claim 14. *$d < n'/2$.*

Proof. Suppose $d \ge n'/2$, then by the Plotkin bound stated in Theorem 13, D has at most $2n' = 2\epsilon n < n$ codewords, a contradiction. ■

5 Proof Sketch of the Main Theorem

In this section, we give a proof sketch of our main theorem.

Theorem 1 (Main Theorem)**.** *Let $\Delta(\gamma) = \frac{5\gamma}{8} - \frac{\gamma^2}{32}$. For all ϵ, $1/4 \le \epsilon \le 1/2$ and for all γ, $0 < \gamma \le 1$,*

$$\textsc{rej}(\epsilon) \ge \epsilon + \min\{4096(1-\Delta(\gamma))^3\epsilon^3(1-2\epsilon)^{12}, \frac{\gamma}{2}\epsilon(1-2\epsilon)^4\}.$$

Our proof will rely on the following coding theorem which bounds the number of codewords of weight at least w. This is a slightly stronger variant of the well-known Johnson bound, for a proof see, e.g., the Appendix in [2].

Theorem 15 (Johnson bound)**.** *Let C be a binary code of block length n and minimum distance d. Let $B'(n,d,w)$ denote the maximum number of codewords in C of weight* at least *w, then $B'(n,d,w) \le \frac{nd}{nd-2w(n-w)}$.*

The basic idea of the proof is as follows. Since there is a one-to-one correspondence between the weight of codeword in $C+v$ and that of D, we will be working with the spectrum of D. Since D is a linear code, its minimum distance d is equal to the minimum weight of its codewords. If d is small (much smaller than $n'/2$), then there is low weight codeword in D. Consequently, there is an $x_i = -\delta n$ for some positive δ, which implies a large gain by Lemma 8. However, if d is large (very close to $n'/2$), then we can apply the Johnson bound to D to show that the number of x_i such that $x_i \geq (1-2\epsilon-\eta)n$ is less than $\frac{1}{(1-2\epsilon)^2}$ for some positive η. This also implies a large gain by Lemma 9. Moreover, as shown below in Lemma 16, there is a trade-off relation between these two gains: If δ is small then η is large and vice versa. This trade-off enables us to prove that $\textsc{gain}(\epsilon) = \Omega(1)$ for every ϵ, $1/4 \leq \epsilon < 1/2$.

Proof sketch of the Main Theorem. By Lemma 11, for all $i > 0$, $4\mathrm{wt}(d_i) - 2\epsilon n \leq (1-2\epsilon)n$. Since for all $i > 0$, $x_i \leq (1-2\epsilon)n$, it follows that $\mathrm{wt}(d_i) \leq \frac{n}{4} = \frac{n'}{4\epsilon}$. Suppose the minimum distance of D is $d = (1/2-\delta')n'$. By Claim 14, δ' is positive.

Note that $x_0 = (1-2\epsilon)n$ and for all $i > 0$, $x_i \geq (1-2\epsilon-\eta)n$ *iff* $\mathrm{wt}(d_i) \geq (\frac{1}{4\epsilon} - \eta')n'$, where $\eta' = \frac{\eta}{4\epsilon}$. Therefore, in order to apply Lemma 9, it suffices to show that there are at most $\frac{1}{(1-2\epsilon)^2} - 2$ codewords in D of weight at least $w' \stackrel{\text{def}}{=} (\frac{1}{4\epsilon} - \eta')n'$ for some $\eta' > 0$. In the next lemma, we show a trade-off relation between δ' and η'. The proof of this lemma appears in Appendix C.

Lemma 16 (Trade-off Lemma)**.** *For every ϵ, $\frac{1}{4} \leq \epsilon < \frac{1}{2}$, there exist two positive numbers δ_0 and η_0 which depend only on ϵ, and a function f which is parameterized only by ϵ and is monotone decreasing in $[0, \eta_0]$, such that the following holds: For all δ' with $0 < \delta' < \delta_0$, let $\eta' = f(\delta')$, then if the minimum distance of code D is $(\frac{1}{2} - \delta')n'$, then D has at most $\frac{1}{(1-2\epsilon)^2} - 2$ codewords of weight at least $(\frac{1}{4\epsilon} - \eta')n'$.*

Combine this Trade-off Lemma with the two lemmas regarding $\textsc{gain}(\epsilon)$, Lemma 8 and Lemma 9, we get the following lower bound for $\textsc{gain}(\epsilon)$. The proof of this lemma appears in Appendix D. Note that since f is monotone in $[0, \eta_0]$, the inverse of f exists in this interval, which we denote by f^{-1}.

Lemma 17 (Gain Lemma)**.** *For* all $\eta' \in (0, \eta_0)$, *let* $\delta' = f^{-1}(\eta')$, *then* $\textsc{gain}(\epsilon) \geq \min\{128(\epsilon\delta')^3, 4\epsilon(1-2\epsilon)^2\eta'\}$.

By choosing η' appropriately and plugging η' and $\delta' = f^{-1}(\eta')$ into Lemma 17, we can show that

$$\textsc{gain}(\epsilon) \geq \min\{8192(1-\Delta(\gamma))^3\epsilon^3(1-2\epsilon)^{12}, \gamma\epsilon(1-2\epsilon)^4\}.$$

Then the Main Theorem follows. Details can be found in Appendix D. ■

Acknowledgment N.X. is very grateful to Ronitt Rubinfeld for making his visit to MIT possible. We thank Ronitt Rubinfeld, Madhu Sudan and Luca Trevisan for helpful discussions.

References

[1] M. Bellare, D. Coppersmith, J. Håstad, M. Kiwi, and M. Sudan. Linearity testing over characteristic two. *IEEE Transactions on Information Theory*, 42(6):1781–1795, 1996. Earlier version in FOCS'95.

[2] M. Bellare, O. Goldreich, and M. Sudan. Free bits, PCP and non-approximability - towards tight results. *SIAM J. on Comput.*, 27(3):804–915, 1998. Earlier version in FOCS'95.

[3] M. Bellare, S. Goldwasser, C. Lund, and A. Russell. Efficient probabilistically checkable proofs and applications to approximation. In *Proc. 25th Annual ACM Symposium on the Theory of Computing*, pages 304–294, 1993.

[4] M. Bellare and M. Sudan. Improved non-approximability results. In *Proc. 26th Annual ACM Symposium on the Theory of Computing*, pages 184–193, 1994.

[5] M. Blum, M. Luby, and R. Rubinfeld. Self-testing/correcting with applications to numerical problems. *J. Comp. Sys. Sci.*, 47:549–595, 1993. Earlier version in STOC'90.

[6] T. Gowers. A new proof of Szemerédi's theorem. *Geometric and Functional Analysis*, 11(3):465–588, 2001.

[7] J. Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798–859, 2001. Earlier version in STOC'97.

[8] T. Kaufman and S. Litsyn. Almost orthogonal linear codes are locally testable. In *Proc. 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 317–326, 2005.

[9] M. Kiwi. Algebraic testing and weight distributions of codes. *Theor. Comp. Sci.*, 299(1-3):81–106, 2003. Earlier version appeared as ECCC TR97-010, 1997.

[10] M. Plotkin. Binary codes with specified minimum distance. *IRE Transactions on Information Theory*, 6:445–450, 1960.

[11] K.F. Roth. On certain sets of integers. *J. London. Math. Soc.*, 28:245–252, 1953.

[12] A. Samorodnitsky and L. Trevisan. A PCP characterization of NP with optimal amortized query complexity. In *Proc. 32nd Annual ACM Symposium on the Theory of Computing*, pages 191–199, 2000.

[13] A. Samorodnitsky and L. Trevisan. Gower uniformity, influence of variables and PCPs. In *Proc. 38th Annual ACM Symposium on the Theory of Computing*, pages 11–20, 2006.

[14] E. Szemerédi. On sets of integers containing no k elements in arithmetic progression. *Acta Arith.*, 27:299–345, 1975.

[15] L. Trevisan. Recycling queries in PCPs and linearity tests. In *Proc. 30th Annual ACM Symposium on the Theory of Computing*, pages 299–308, 1998.

A Two inequalities

We will use the following two inequalities in our analysis.

Lemma 18. *For all real y with $0 \leq y \leq 1/2$,*

$$\frac{1}{1-y} - y \geq \frac{1}{\sqrt{1-2y^2}}.$$

Proof. By Taylor expansion, we have

$$\frac{1}{1-y} - y = 1 + \sum_{k=2}^{\infty} y^k,$$

and

$$\frac{1}{\sqrt{1-2y^2}} = 1 + \sum_{k=1}^{\infty} \frac{(2k-1)!!}{k!} y^{2k}.$$

It is easy to show by induction that for all $k \geq 2$, $\frac{(2k-1)!!}{k!2^{k-1}} < 1$. Indeed, $\frac{1\cdot 3}{2!2} = \frac{3}{4} < 1$, and $\frac{(2k+1)!!}{(k+1)!2^k} = \frac{(2k-1)!!}{k!2^{k-1}} \frac{2k+1}{2k+2} < \frac{(2k-1)!!}{k!2^{k-1}} < 1$, where the last inequality follows from induction hypothesis. Hence we have

$$\begin{aligned}
\frac{1}{\sqrt{1-2y^2}} &= 1 + y^2 + \sum_{k=2}^{\infty} \frac{(2k-1)!!}{k!} y^{2k} \\
&\leq 1 + y^2 + \sum_{k=2}^{\infty} \frac{(2k-1)!!}{k!} \frac{1}{2^{k-1}} y^{k+1} \quad \text{(because } y \leq 1/2\text{)} \\
&\leq 1 + y^2 + \sum_{k=2}^{\infty} y^{k+1} \\
&= \frac{1}{1-y} - y.
\end{aligned}$$

■

Lemma 19. *Let γ be a constant with $0 \leq \gamma \leq 1$. Then for all real y with $0 \leq y \leq 1/2$,*

$$\frac{1}{(1-y)^2} - \frac{1}{1-2y^2} - \gamma \frac{y}{1-y} \geq (8 - 5\gamma) y^2.$$

Proof. We break the proof into two parts: First we show the inequality holds for $0 \leq y \leq 2/7$, then we prove it for the interval $2/7 \leq y \leq 1/2$.

Proposition 20. *For all y and γ with $0 \leq y \leq \frac{2}{7}$ and $0 \leq \gamma \leq 1$,*

$$\frac{1}{(1-y)^2} - \frac{1}{1-2y^2} - \gamma \frac{y}{1-y} \geq (8 - 5\gamma) y^2.$$

Proof. By Taylor expansion,

$$\begin{aligned}
&\frac{1}{(1-y)^2} - \frac{1}{1-2y^2} - \gamma \frac{y}{1-y} - (8-5\gamma)y^2 \\
=&\sum_{k=0}^{\infty} (k+1) y^k - \sum_{k=0}^{\infty} (2y^2)^k - \gamma \sum_{k=1}^{\infty} y^k - (8-5\gamma)y^2 \\
=&(2-\gamma)y - (7-4\gamma)y^2 + (4-\gamma)y^3 + (1-\gamma)y^4 + \sum_{k=5}^{\infty} (k+1-\gamma) y^k - \sum_{k=3}^{\infty} (2y^2)^k \\
\geq&(2-\gamma)y - (7-4\gamma)y^2 + (4-\gamma)y^3 + (1-\gamma)y^4 - \frac{8y^6}{1-2y^2} \\
\geq&(2-\gamma)y - (7-4\gamma)y^2 + 3y^3 - \frac{8y^6}{1-2y^2}.
\end{aligned}$$

Since $0 \le y \le \frac{2}{7}$, $(2-\gamma)y \ge \frac{7}{2}(2-\gamma)y^2 = (7-\frac{7}{2}\gamma)y^2 \ge (7-4\gamma)y^2$, $\frac{8y^6}{1-2y^2} \le \frac{8y^6}{1-2(\frac{2}{7})^2} \le 10y^6$, and $3y^3 \ge 3(\frac{7}{2})^3 y^6 \ge 10y^6$, this completes the proof of the Proposition. ■

Proposition 21. *For all y and γ with $\frac{2}{7} \le y \le \frac{1}{2}$ and $0 \le \gamma \le 1$,*

$$\frac{1}{(1-y)^2} - \frac{1}{1-2y^2} - \gamma\frac{y}{1-y} \ge (8-5\gamma)y^2.$$

Proof. Let $z = 1-2y$. After substituting z into the expression and some simplifications, we see that proving the original inequality is equivalent to proving, for $0 \le z \le \frac{3}{7}$,

$$\frac{4}{(1+z)^2} - \frac{2}{2-(1-z)^2} - \gamma\frac{1-z}{1+z} \ge (2-\frac{5}{4}\gamma)(1-z)^2.$$

Or, after dividing $(1-z)^2$ on both sides,

$$\frac{4}{(1-z^2)^2} - \frac{2}{2(1-z)^2-(1-z)^4} - \gamma\frac{1}{1-z^2} \ge 2 - \frac{5}{4}\gamma.$$

Note that since $0 \le z \le \frac{3}{7}$, we have $\frac{\gamma}{1-z^2} \le \frac{\gamma}{1-(\frac{3}{7})^2} = \frac{49}{40}\gamma \le \frac{5}{4}\gamma$, so the only thing remains to show is $\frac{4}{(1-z^2)^2} - \frac{2}{2(1-z)^2-(1-z)^4} \ge 2$. Indeed,

$$\begin{aligned}
&\frac{4}{(1-z^2)^2} - \frac{2}{2(1-z)^2-(1-z)^4} \ge 2 \\
\iff &\frac{4}{(1-z^2)^2} - \frac{2}{1-(2z-z^2)^2} \ge 2 \\
\iff &\frac{2}{(1-z^2)^2} \ge \frac{2-z^2(2-z)^2}{1-z^2(2-z)^2} \\
\iff &2(1-z^2(2-z)^2) \ge (1-z^2)^2(2-z^2(2-z)^2) \\
\iff &2(1+2z-z^2) \ge 2(1+z)^2 - z^2(1+z)^2(2-z)^2 \\
\iff &(1+z)^2(2-z)^2 \ge 4 \\
\iff &z(1-z) \ge 0.
\end{aligned}$$

This finishes the proof of the Proposition. ■

This completes the proof of Lemma 19. ■

B Proofs of Lemma 7

Lemma 7. *Consider the following Maximal Sum of Cubes Problem: For a set of n integers $x_0, x_1, \cdots, x_{n-1}$, find the maximum of $x_0^3 + x_1^3 + \cdots + x_{n-1}^3$ under the constraints:*

$$x_0^2 + x_1^2 + \cdots + x_{n-1}^2 = n^2$$
$$\forall i : x_i \le \alpha n, \text{ with } \alpha \le 1.$$

The maximum is achieved when $\frac{1}{\alpha^2}$ of the x_i's get the maximum value αn, and the rest are zeros. The maximum thus obtained is αn^3.

Proof. Indeed,

$$\frac{(\sum_{j=0}^{n-1} x_j^3)^{1/3}}{(\sum_{j=0}^{n-1} x_j^2)^{1/2}} = (\sum_{j=0}^{n-1} \frac{x_j^3}{(\sum_{i=0}^{n-1} x_i^2)^{3/2}})^{1/3} = (\sum_{j=0}^{n-1} (\frac{x_j^2}{\sum_{i=0}^{n-1} x_i^2})^{3/2})^{1/3}.$$

Now notice that since $0 \le x_i^2 \le \alpha^2 n^2$,

$$(\frac{x_j^2}{\sum_{i=0}^{n-1} x_i^2})^{3/2} \le \alpha \frac{x_j^2}{\sum_{i=0}^{n-1} x_i^2}.$$

Therefore,

$$\frac{(\sum_{j=0}^{n-1} x_j^3)^{1/3}}{(\sum_{j=0}^{n-1} x_j^2)^{1/2}} \le (\alpha \sum_{j=0}^{n-1} \frac{x_j^2}{\sum_{i=0}^{n-1} x_i^2})^{1/3} = (\alpha)^{1/3}.$$

Moreover, the equality is attained only if all of the values of x_i are either zero or αn. This is possible only if $\frac{1}{\alpha^2}$ of the x_i's equal αn, and the rest are zeros. In that case $x_0^3 + x_1^3 + \cdots + x_{n-1}^3 = \alpha n^3$. ∎

C Proof on the Trade-off Lemma

In this section we prove the Trade-off Lemma (Lemma 16).

Proof. Let $\eta' = f(\delta')$. We apply the Johnson bound stated in Theorem 15 to code D. Plug in $d = (1/2 - \delta')n'$ and $w' = (\frac{1}{4\epsilon} - \eta')n'$, we get

$$\frac{\frac{1}{2} - \delta'}{(\frac{1}{2} - \delta') - 2(\frac{1}{4\epsilon} - \eta')(1 - \frac{1}{4\epsilon} + \eta')} = \frac{1}{(1-2\epsilon)^2} - 2. \tag{1}$$

If we solve (1) to get $f(\delta') = \eta'$, then the statement in the lemma about η' is also true for all $\eta'' \le \eta'$, provided η' is not too large[2]. By some elementary algebraic manipulations, we have

$$\begin{aligned}\delta' &= \frac{1}{2} - 2(\frac{1}{4\epsilon} - \eta')(1 - \frac{1}{4\epsilon} + \eta')\frac{1 - 2(1-2\epsilon)^2}{1 - 3(1-2\epsilon)^2} \\ &= \frac{2(1 - 2(1-2\epsilon)^2)}{1 - 3(1-2\epsilon)^2} g(\eta'),\end{aligned}$$

where $g(\eta') \stackrel{\text{def}}{=} (\eta' - \frac{1}{4\epsilon})(\eta' - \frac{1}{4\epsilon} + 1) + \frac{1-3(1-2\epsilon)^2}{4(1-2(1-2\epsilon)^2)}$. Note that since $1/4 \le \epsilon < 1/2$, we have both $1 - 2(1-2\epsilon)^2$ and $1 - 3(1-2\epsilon)^2$ are positive. Therefore, whenever there are positive values η' to make $g(\eta')$ positive, the corresponding δ' will be positive as well.

Rewrite $g(\eta')$ as $g(\eta') = \eta'^2 - b\eta' + c$, where $b = \frac{1}{2\epsilon} - 1 > 0$ and $c = \frac{1-3(1-2\epsilon)^2}{4(1-2(1-2\epsilon)^2)} - \frac{1}{4\epsilon} + \frac{1}{16\epsilon^2}$. Since $b^2 - 4c = \frac{(1-2\epsilon)^2}{1-2(1-2\epsilon)^2} > 0$, there are two real roots for $g(\eta') = 0$. Denote these two roots by η_1

[2]That is, we require that $x \stackrel{\text{def}}{=} \frac{1}{4\epsilon} - \eta' > \frac{1}{2}$. Since the function $x(1-x)$ is monotone decreasing for $\frac{1}{2} < x < 1$, plugging some $\eta'' < \eta'$ into (1) will only make the LHS smaller thus changing the equality into an inequality.

and η_2 with $\eta_1 > \eta_2$. Then $g(\eta')$ assumes positive values for $\eta' > \eta_1$ and $\eta' < \eta_2$. Since $\eta_1 > \frac{1}{4\epsilon} - \frac{1}{2}$ but we are bounding the number of codewords of weight at least $w' = (\frac{1}{4\epsilon} - \eta')n' > \frac{1}{2}n'$, which requires $\eta' < \frac{1}{4\epsilon} - \frac{1}{2}$, so we only need to look at the region where $\eta' < \eta_2$. Therefore, we have:

$$\begin{aligned}
&\text{There are positive } \eta' \text{ to make } g(\eta') \text{ positive} \\
\iff &\eta_2 > 0 \\
\iff &c > 0 \\
\iff &\frac{1-3(1-2\epsilon)^2}{4(1-2(1-2\epsilon)^2)} - \frac{1}{4\epsilon} + \frac{1}{16\epsilon^2} > 0 \\
\iff &(\frac{1}{2\epsilon} - 1)^2 > \frac{(1-2\epsilon)^2}{1-2(1-2\epsilon)^2} \\
\iff &\epsilon > 1/6.
\end{aligned}$$

I.e., for all ϵ, $1/4 \le \epsilon < 1/2$, $\eta_2 > 0$. Note that $g(\eta')$ is monotone decreasing in $[0, \eta_2]$, so the inverse of g exists, which we denote by g^{-1}. Finally, we set $f(\delta') = \frac{1-3(1-2\epsilon)^2}{2(1-2(1-2\epsilon)^2)} g^{-1}(\delta')$, $\delta_0 = \frac{2(1-2(1-2\epsilon)^2)}{1-3(1-2\epsilon)^2} c$, and $\eta_0 = \eta_2$ to complete the proof. ∎

D Missing details in the proof of the Main Theorem

In this section we prove the Gain Lemma (Lemma 17) and finish the proof of the Main Theorem.

Lemma 17. *For* all $\eta' \in (0, \eta_0)$, *let* $\delta' = f^{-1}(\eta') = \frac{2(1-2(1-2\epsilon)^2)}{1-3(1-2\epsilon)^2} g(\eta')$, *then* $\text{GAIN}(\epsilon) \ge \min\{128(\epsilon\delta')^3, 4\epsilon(1-2\epsilon)^2\eta'\}$.

Proof. As before, we set $\delta = 4\epsilon\delta'$ and $\eta = 4\epsilon\eta'$. In the following, we consider ϵ to be any fixed value in $[\frac{1}{4}, \frac{1}{2})$. Suppose the minimum distance of D is $(\frac{1}{2} - \delta')n'$. Then on the one hand, there is an x_i, such that $x_i = -4\epsilon\delta' n = -\delta n$. On the other hand, by Lemma 16, there are at most $\frac{1}{(1-2\epsilon)^2} - 2$ codewords of weight at least $(\frac{1}{4\epsilon} - \eta')n'$ in D, which implies that there are at most $\frac{1}{(1-2\epsilon)^2} - 1$ x_i's such that $x_i \ge (1 - 2\epsilon - 4\epsilon\eta')n = (1 - 2\epsilon - \eta)n$. Denote the gains as functions of η' given in Lemma 8 and Lemma 9 by $\text{GAIN}_\delta(\eta')$ and $\text{GAIN}_\eta(\eta')$, respectively. Then we have $\text{GAIN}_\delta(\eta') = 2\delta^3 = 128\epsilon^3\delta'^3$ and $\text{GAIN}_\eta(\eta') = (1-2\epsilon)^2\eta = 4\epsilon(1-2\epsilon)^2\eta'$. Therefore $\text{GAIN}(\epsilon) \ge \min_{0<\eta'<\eta_0} \max\{\text{GAIN}_\delta(\eta'), \text{GAIN}_\eta(\eta')\}$. Note that GAIN_δ is monotone increasing in δ' and GAIN_η is monotone increasing in η'. At one end $\eta' = 0$, $\text{GAIN}_\delta > 0$ and $\text{GAIN}_\eta = 0$; at the other end $\eta' = \eta_0$, $\text{GAIN}_\delta = 0$ and $\text{GAIN}_\eta > 0$. Combine these with the fact that $g(\eta')$ is monotone decreasing, we see that there exists an η'', $0 < \eta'' < \eta_0$, such that $\text{GAIN}_\delta(\eta'') = \text{GAIN}_\eta(\eta'') = \min_{0<\eta'<\eta_0} \max\{\text{GAIN}_\delta(\eta'), \text{GAIN}_\eta(\eta')\}$. By monotonicity again, for all $\eta' \in (0, \eta_0)$, $\text{GAIN}(\epsilon) \ge \min\{\text{GAIN}_\delta(\eta'), \text{GAIN}_\eta(\eta')\}$. ∎

Next we seek an explicit bound of $\text{GAIN}(\epsilon)$. We begin with a simple lower bound for η_0.

Claim 22.

$$\eta_0 \ge \frac{(1-2\epsilon)^2}{2}.$$

Proof. By definition,

$$\eta_0 = \eta_2 = \frac{1}{2}\left(\frac{1}{2\epsilon} - 1 - \sqrt{\frac{(1-2\epsilon)^2}{1-2(1-2\epsilon)^2}}\right) = \frac{1-2\epsilon}{2}\left(\frac{1}{2\epsilon} - \frac{1}{\sqrt{1-2(1-2\epsilon)^2}}\right).$$

Now change the variable from ϵ to $y = 1-2\epsilon$ and apply Lemma 18, the desired bound follows. ■

Set $\eta' = \gamma\frac{(1-2\epsilon)^2}{4}$, where $0 < \gamma \leq 1$ is a constant. Plugging η' into $g(\eta')$ and after some straightforward calculations, we get

$$g(\eta') = \frac{(1-2\epsilon)^2}{4}\left(\frac{\gamma^2}{4}(1-2\epsilon)^2 - \gamma\frac{1-2\epsilon}{2\epsilon} + \frac{1}{4\epsilon^2} - \frac{1}{1-2(1-2\epsilon)^2}\right).$$

By changing variable to $y = 1-2\epsilon$ and applying Lemma 19, we arrive at

$$\begin{aligned} g(\eta') =& \frac{y^2}{4}\left(\frac{\gamma^2}{4}y^2 - \gamma\frac{y}{1-y} + \frac{1}{(1-y)^2} - \frac{1}{1-2y^2}\right) \\ \geq& \frac{y^2}{4}\left(\frac{\gamma^2}{4}y^2 + (8-5\gamma)y^2\right) \\ =& 2(1-\Delta(\gamma))(1-2\epsilon)^4, \end{aligned}$$

where $\Delta(\gamma) \stackrel{\text{def}}{=} \frac{5\gamma}{8} - \frac{\gamma^2}{32}$. Therefore,

$$\delta' = \frac{2(1-2(1-2\epsilon)^2)}{1-3(1-2\epsilon)^2}g(\eta') \geq 2g(\eta') \geq 4(1-\Delta(\gamma))(1-2\epsilon)^4.$$

Plugging η' and δ' into Lemma 17, we get

$$\textsc{gain}(\epsilon) \geq \min\{8192(1-\Delta(\gamma))^3\epsilon^3(1-2\epsilon)^{12}, \gamma\epsilon(1-2\epsilon)^4\}.$$

This completes the proof of the Main Theorem.

ECCC ISSN 1433-8092

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