New Direct Sum Tests

Alek Westover
MIT, USA

Edward Yu
MIT, USA

Kai Zhe Zheng
MIT, USA

Abstract

A function $f : [n]^d \to \mathbb{F}_2$ is a ***direct sum*** if there are functions $L_i : [n] \to \mathbb{F}_2$ such that $f(x) = \sum_i L_i(x_i)$. In this work we give multiple results related to the property testing of direct sums.

Our first result concerns a test proposed by Dinur and Golubev in [13]. We call their test the Diamond test and show that it is indeed a direct sum tester. More specifically, we show that if a function f is ε-far from being a direct sum function, then the Diamond test rejects f with probability at least $\Omega_{n,\varepsilon}(1)$. Even in the case of $n = 2$, the Diamond test is, to the best of our knowledge, novel and yields a new tester for the classic property of affinity.

Apart from the Diamond test, we also analyze a broad family of direct sum tests, which at a high level, run an arbitrary affinity test on the restriction of f to a random hypercube inside of $[n]^d$. This family of tests includes the direct sum test analyzed in [13], but does not include the Diamond test. As an application of our result, we obtain a direct sum test which works in the online adversary model of [20].

Finally, we also discuss a Fourier analytic interpretation of the diamond tester in the $n = 2$ case, as well as prove local correction results for direct sum as conjectured by [13].

2012 ACM Subject Classification Theory of computation → Computational complexity and cryptography

Keywords and phrases Linearity testing, Direct sum, Grids

Digital Object Identifier 10.4230/LIPIcs.ITCS.2025.20

Related Version arxiv
arxiv: https://arxiv.org/abs/2408.16092

Funding *Alek Westover*: Supported by MIT SPUR
Edward Yu: Supported by MIT SPUR
Kai Zhe Zheng: Department of Mathematics, Massachusetts Institute of Technology, Cambridge, USA. Supported by the NSF GRFP DGE-2141064

1 Introduction

In property testing, one is given query access to a function over a large, typically multidimensional domain, say $f : \prod_{i=1}^{d} S_i \to T$. The goal is to determine whether or not f satisfies some property, $\mathcal{P}$, using as few queries to f as possible. Here, a property $\mathcal{P}$ can be any subset of functions. It is not hard to see that distinguishing between $f \in \mathcal{P}$ and $f \notin \mathcal{P}$ requires querying the entire domain for any nontrivial property $\mathcal{P}$, and so we typically settle for an algorithm with a weaker guarantee, typically called a property tester. A tester for $\mathcal{P}$ is a randomized algorithm, which given input function f, makes queries to f and satisfies the following two properties:

- **Completeness**: If $f \in \mathcal{P}$, the tester accepts with probability 1.
- **Soundness**: If f is δ-far from $\mathcal{P}$, the tester rejects with probability $s(\delta)$.

We say that f is δ-far from $\mathcal{P}$, if the minimal fractional hamming distance between f and any $g \in \mathcal{P}$ is at least δ. The function $s(\delta)$ is referred to as the soundness of the tester and oftentimes one repeats the tester in $s(\delta)^{-1}$ times to reject with probability 2/3 in the soundness case. It is clear that for any property, there is a trivial algorithm that queries f on its entire domain, so oftentimes the goal is to design testers whose query complexity is sublinear in the domain size or even independent of the dimension d.

In this paper we consider two property testing questions related to functions known as direct sums. A function f is called a direct sum if it can be written as a sum of functions on the individual coordinates, that is

$$f(x_1, \dots, x_d) = \sum_{i=1}^{d} L_i(x_i),$$

for some functions $L_i : S_i \to T$. Direct sums are a natural property to consider in the context of property testing. Indeed, when one takes $f : \mathbb{F}_2^d \to \mathbb{F}_2$, then the direct sum property is equivalent to the classic property of affinity (being a linear function plus a constant) considered by Blum, Luby, and Rubinfeld's seminal work [8]. Property testing of more general versions of direct sums, called low junta-degree functions, has also been considered before in the works of [5, 3]. A junta-degree t function is one that can be written as the sum of t-juntas. Thus, direct sums are junta-degree 1 functions, and direct sum testing is a specialization of the low junta-degree testing problem that appears in [5, 3].

In addition to being a natural property to study, motivation for direct sum testing also comes from its relation to the more widely known direct product testing problem. Historically, direct product testing was first introduced by Goldreich and Safra in [17] due to its potential for application as a hardness amplification step in more efficient PCP constructions [15, 12, 14, 18, 4]. More recently, the related problem of direct sum testing was considered, first by [10] and later by [13]. The motivation for considering direct sums stems in part from the fact that direct sums are similar to direct products in that they can be used to amplify hardness, but with the advantage that the output is a single value, as opposed to an entire tuple. This seems to make direct sum testing more difficult, but has the potential for leading to even more efficient PCPs (in particular, in improving a parameter known as the alphabet size). Analysis of a type of direct sum test is a key piece of Chan's elegant PCP construction in [9].

Direct Sum Testing In [13], Dinur and Golubev give and analyze a direct sum tester which they call the Square in a Cube test. They propose an additional test which they refer to as the diamond test, but leave its analysis for future work. We call this second test the Diamond Test, and describe both of these tests below.

Let the domain be an arbitrary grid $[\overline{n}; d] = \prod_{i=1}^{d}[n_i]$ where each $[n_i] = \{1, \ldots, n_i\}$. We will let the output be $\mathbb{F}_2$ so that our input function is

$$f : [\overline{n}; d] \to \mathbb{F}_2.$$

Given two points $a, b \in [\overline{n}; d]$ and $x \in \{0, 1\}^d$, let ***interpolation*** $\phi_x(a, b)$ to be the vector obtained by taking a_i whenever $x_i = 0$ and b_i whenever $x_i = 1$. The Square in a Cube test proceeds as follows.

▶ **Test 1** (Square in a Cube)**.** Sample random $a, b \in [\overline{n}; d]$ and $x, y \in \{0, 1\}^d$. Accept if and only if

$$f(a) + f(\phi_x(a, b)) + f(\phi_y(a, b)) + f(\phi_{x+y}(a, b)) = 0.$$

The Diamond Test proceeds in a slightly different manner, which essentially fixes $x + y = (1, \ldots, 1)$.

▶ **Test 2** (Diamond)**.** Sample random $a, b \in [\overline{n}; d], x \in \{0, 1\}^d$. Accept iff

$$f(a) + f(\phi_x(a, b)) + f(\phi_x(b, a)) + f(b) = 0.$$

In addition to the Diamond test, we also consider a family of tests that generalizes the Square in a Cube test. We call this test the Affine on Subcube test.

▶ **Test 3** (Affine on Subcube)**.** Sample $a, b \in [n]^d$ and run an affinity test on the function $x \mapsto f(\phi_x(a, b))$.

Note that the affinity test can be any arbitrary affinity test for functions $\mathbb{F}_2^d \to \mathbb{F}_2$. The Square in a Cube test is a special case of the above with the classic BLR affinity test as the affinity test. The Diamond Test, however, is not a specialization of the above.

As is usual in property testing, it is clear that both the Diamond and Affine on Subcube test satisfy completeness, so our main interest is with regards to their soundness.

1.1 Main Results

We now describe our main results. Our first result establishes soundness for the Diamond test. As the completeness is clear, this shows that the Diamond test is indeed a Direct Sum tester.

▶ **Theorem 4.** *Suppose $f : [n]^d \to \mathbb{F}_2$ passes the Diamond test with probability $1 - \varepsilon$. Then, f is $C_n \cdot \varepsilon$-close to a direct sum, for some constant C_n independent of d.*

We remark that the theorem above incurs a dependence on n. In contrast, [13] shows a version of the above result with C_n replaced by an absolute constant, C, independent of n. For $n = O(1)$, the soundness we show for the Diamond Test is comparable to what is known for the Square in a Cube test, but for larger n the soundness analysis becomes weaker. This dependence on n in the soundness is not so uncommon in property testing results however. It is comparable to the dependence on field size in the low degree testing results of [2, 21], or on grid size in the junta degree testing results of [5, 3]. We leave it to future work to remove the dependence on n in the soundness.

Our next result is a reformulation of Theorem 4 in the $n = 2$ case. Here the Diamond Test is, to the best of our knowledge, a novel affinity tester. Moreover, we observe that the soundness of the $n = 2$ Diamond Test is equivalent to a succinct Fourier analytic fact regarding a function's distance to affinity, or equivalently its maximum magnitude Fourier coefficient. Interestingly, we were unable to show this fact without appealing to the soundness of the Diamond test. Intuitively, the Fourier analytic fact says "f is close to affine if and only if when you randomly fix or don't fix each coordinate of f to a random value, f becomes close to an even or odd function". Formally, the result is:

▶ **Theorem 5.** *For $f : \mathbb{F}_2^d \to \mathbb{F}_2$, define $\mathcal{R}(f)$ to be a random restriction of f obtained as follows: for each $i \in [d]$ independently, with probability $1/4$ restrict x_i to 0, with probability $1/4$ restrict x_i to 1, and with probability $1/2$ do not restrict x_i. Then, letting* EvenOrOdd *denote the set of functions that are either even or odd, we have:*

$$\mathsf{dist}(f, \mathsf{affine}) = \Theta(\mathbb{E}_{\mathcal{R}}[\mathsf{dist}(\mathcal{R}(f), \mathsf{EvenOrOdd})]).$$

Along with the Diamond Test, we also describe and analyze a family of testers for direct sum which we call the Affine on Subcube test (Test 3).

▶ **Theorem 6.** *Fix $f : [n]^d \to \mathbb{F}_2, \varepsilon \geq 0$ and an affinity test T. If f passes the Affine on Subcube test (instantiated with T) with probability $1 - \varepsilon$, then f is $C\varepsilon$-close to a direct sum, for some constant C independent of n, d.*

By using a suitable erasure-resilient tester as the inner affinity tester in Test 3, we use Theorem 36 obtain direct sum testers in the online-erasure model recently introduced by [20]. We discuss this result as well as the online-erasure model in more detail in Section 4.2.

Finally, we address a second question posed in [13] regarding reconstructing functions which are close to direct sums. They ask if a test that they call the Shapka test can be used to reconstruct, via a "voting scheme", a direct sum given query access to its corrupted version. We show that this is indeed the case, and also give an improved reconstruction method that uses fewer queries.

▶ **Proposition 7.** *Let $f : [n]^d \to \mathbb{F}_2$ be ε close to a direct sum L, with $(n+1)\varepsilon < 1/4$. Then L can be reconstructed using a voting scheme on f using n queries to f.*

We also give lower bounds on the query complexity needed and show that our improved method is asymptotically optimal (up to constant factors) in terms of both query complexity and the fraction of corruptions it can tolerate.

Proof Overview

We give an overview of the proof of Theorem 4. Our analysis here diverges significantly from that of Dinur and Golubev for the Square in a Cube Test. Before discussing our analysis, we first summarize their approach and where it fails for the Diamond Test.

The underlying idea behind the Square in a Cube Test is that after choosing two points $a, b \in [n]^d$, any direct sum restricted to the hypercube-like domain $\prod_{i=1}^d \{a_i, b_i\}$ must be an affine function (i.e multivariate polynomial of degree 1). This idea gives way to a natural interpretation of the Square in a Cube Test

- Choose two points $a, b \in [n]^d$ randomly and view the domain $\prod_{i=1}^d \{a_i, b_i\}$ as a hypercube of the appropriate dimension d'. Here d' is the number of coordinates on which a and b differ.
- Define the function $f_{a,b} : \{0,1\}^{d'} \to \mathbb{F}_2$ by $f_{a,b}(x) = f(\phi_x(a, b))$.

- Perform the BLR affinity test on $f_{a,b}$. That is, choose $x, y, z \in \{0,1\}^{d'}$ and accept if and only if

$$f_{a,b}(x) + f_{a,b}(x+y) + f_{a,b}(x+z) + f_{a,b}(x+y+z) = 0.$$

From here, the analysis of [13] has two parts. First, they apply BLR to show that if the overall Square in a Cube test passes with high probability, then for many a, b, the function $f_{a,b}$ is close to an affine function (or direct sum), say $F_{a,b}$, on the hypercube $\prod_{i=1}^{d}\{a_i, b_i\}$. These $F_{a,b}$ can be thought of as local direct sums that agree with f locally on hypercubes inside of $[n]^d$. The second step is to show that many of these $F_{a,b}$ are actually consistent with each other and apply a direct product testing result of [16] to conclude that there is in fact a global direct sum F consistent with many $F_{a,b}$'s. As the $F_{a,b}$'s are in turn largely consistent with f, we get that F is a direct sum close to f, concluding the proof of soundness.

Our analysis of the soundness of the Diamond test, however, requires a different approach. Let us first see what happens when we try to naively adapt the above strategy. After choosing $a, b \in [n]^d$, instead of applying BLR, the Diamond test performs the following test on $f_{a,b} : \{0,1\}^{d'} \to \mathbb{F}_2$.

Choose $x \in \{0,1\}^{d'}$ and accept if and only if $f_{a,b}(0) + f_{a,b}(x) + f_{a,b}(x + \mathbb{1}) + f_{a,b}(\mathbb{1}) = 0$.

There is an issue though: the above is not an affinity tester! Indeed one can check that if $f_{a,b}(0) + f_{a,b}(\mathbb{1}) = 0$, resp. 1, then any even, resp. odd, function passes the above test with probability 1. Thus, we would only get that many $f_{a,b}$ are close to even or odd functions, and from here it is not clear how to make the second part of the analysis go through.

We instead combine ideas from the soundness analyses of [10, 7, 3], which use induction. Suppose $f : [n]^d \to \mathbb{F}_2$ is ε-far from a direct sum. We wish to show that the Diamond Test rejects f with some non-trivial probability that is independent of d. The inductive approach of both of the mentioned works consist of three main steps: 1) Restate the test on f as choosing a random subdomain that is one dimension and performing the test on f restricted to this subdomain, 2) analyze the soundness in the ε very small case, 3) on the other hand, show that if ε is large then show that f must also be far from a direct sum on many subdomains.

To execute all of the above steps, we have to draw on ideas from all of the mentioned works. For step 1), we take inspiration from [10] and instead analyze a four-function version of the Diamond Test. The reason is that there is no clear way to view the Diamond Test in the manner described above. It is tempting to describe the Diamond Test as choosing a random coordinate, say $i = 1$, and a random value in $[n]$, say 1, and then performing a $d-1$-dimensional version of the Diamond Test on the function $f(1, x_2, \dots, x_d)$. Unfortunately this does not work as one has to change the distribution that the points in the $d-1$-dimensional grid are chosen. Indeed, if one chooses the points $a, b \in [n]^{d-1}$ of the one dimension down Diamond Test uniformly at random, then the final d-dimensional points output would agree on $1 + (d-1)/n$ coordinates in expectation instead of d/n.

For the small distance case in step 2), we rely on the a hypercontractivity theorem over grids. However, since we are now working with a four-function version of the Diamond Test some additional care is needed in this analysis.

Finally for step 3), we use techniques from agreement testing in the PCP literature [25, 23]. From [25] we borrow their transitive-consistency graph technique used to analyze the so-called Plane versus Plane test and from [23] we show a version of their hyperplane sampling lemma for $(d-1)$-dimensional grids in $[n]^d$.

Paper Outline In Section 3 we present and analyze our novel affinity tester (the Diamond test). In Section 4 we give a new class of direct sum tests. In Section 5 we consider the problem of obtaining "corrected" samples from corrupted direct sums.

2 Preliminaries

We now introduce some notations that will be used throughout the paper. We let $[n]$ denote $\{1, 2, \ldots, n\}$. Let $[n; d]$ denote a product space $[n_1] \times \cdots \times [n_d]$. For a set or event X, we write $\overline{X}$ to denote the complement of X.

The notion of distance that we use is fractional hamming distance. That is, given two functions $f : S \to T$ and $g : S \to T$, the distance between them is the fraction of entries on which they differ:

$$\mathsf{dist}(f, g) = \Pr_{x \in S}[f(x) \neq g(x)].$$

Given a property $\mathcal{P} \subseteq \{g : S \to T\}$, the distance from f to the property $\mathcal{P}$ is defined as

$$\mathsf{dist}(f, \mathcal{P}) = \min_{g \in \mathcal{P}} \mathsf{dist}(f, g).$$

We say that f is δ*-far* from $\mathcal{P}$ if $\mathsf{dist}(f, \mathcal{P}) \geq \delta$.

Finally, if we do not explicitly specify the distribution a random variable is drawn from is a probability, then the random variable is meant to be drawn uniformly from the appropriate space (which will be clear from context). The constants hidden in our O-notation are never allowed to depend on d. If there is a parameter k that we think of as constant, we will sometimes use the notation $O_k(\cdot)$ to emphasize that the constant hidden by the O is allowed to depend on k.

Boolean Functions Define the ***interpolation*** $\phi_x(a, b)$ to be the vector obtained by taking a_i whenever $x_i = 0$ and b_i whenever $x_i = 1$. A frequently useful property of interpolations is:

▶ **Fact 8.** $\phi_x(a, b) = \phi_{x+\mathbb{1}}(b, a)$.

Given $x \in \prod_i S_i$ and $p \in [0, 1]$, let $y \sim \mathsf{T}_p(x)$ denote a vector in $\prod_i S_i$, with each y_i sampled independently as follows: set $y_i = x_i$ with probability p, and otherwise sample y_i uniformly from S_i; T_p is called the ***noise operator***. Given $p \in [-1, 0]$ and $x \in \mathbb{F}_2^d$, we define the distribution $y \sim \mathsf{T}_p(x)$ as: $y_i = x_i + 1$ with probability $-p$ and y_i is sampled uniformly from $\{0, 1\}$ otherwise. For a boolean function $f : \mathbb{F}_2^d \to \mathbb{F}_2$ we will use $\mu(f)$ to denote the fractional size of the support of f (i.e., the fraction of $\mathbb{F}_2^d$ on which f outputs 1). We use $\mathbb{1}$ to denote the all ones vector, and sometimes use 0 to denote the zero-vector (the dimension will be clear from context). For $v, w \in \mathbb{F}_2^d$ we write $\langle v, w \rangle$ to denote $\sum_i v_i w_i$. For boolean functions $f, g : \mathbb{F}_2^d \to \mathbb{F}_2$ we write $\langle f, g \rangle$ to denote $\mathbb{E}_x[f(x)g(x)]$.

When discussing functions $f : [n]^d \to \mathbb{F}_2$ we will always assume $n > 1$.

3 The Diamond Test

We restate the Diamond test, initially proposed by Dinur and Golubev in [13] (recall from Section 2 that ϕ_x is the interpolation function)

▶ **Test 9** (Diamond)**.** Sample random $a, b \sim [n]^d, x \sim \mathbb{F}_2^d$. Accept iff

$$f(a) + f(b) = f(\phi_x(a, b)) + f(\phi_x(b, a)).$$

We will show that Test 9 is a valid tester for direct sum. First, we analyze the case that the test passes with probability 1.

▶ **Lemma 10.** *The function $f : [n]^d \to \mathbb{F}_2$ is a direct sum if and only if f passes the Diamond test with probability* 1.

Proof. If f is a direct sum then we can write $f(x) = \sum_i L_i(x_i)$. It follows that for any $a, b \in [n]^d$,

$$f(a) + f(b) = \sum_i (L_i(a_i) + L_i(b_i)) = f(\phi_x(a, b)) + f(\phi_x(b, a)).$$

For the other direction, suppose f passes the Diamond test with probability 1. Then, for all $x \in \mathbb{F}_2^d, a \in [n]^d$ we have

$$f(a) = f(\mathbb{1}) + f(\phi_x(\mathbb{1}, a)) + f(\phi_x(a, \mathbb{1})). \tag{1}$$

Let $e_{i,j} \in [n]^d$ denote a vector which is 1 at all coordinates except coordinate i, where it has value j. Then, repeated application of Equation (1) lets us decompose $f(a)$ as

$$f(a) = f(\mathbb{1}) + \sum_{i=1}^{d} \left(f(e_{i,a_i}) + f(\mathbb{1})\right).$$

Thus, f is a direct sum. ◀

As a simple corollary of Lemma 10 we have:

▶ **Corollary 11.** *If f is ε-close to a direct sum, then f passes the Diamond test with probability at least $1 - 4\varepsilon$.*

Proof. Let g be a direct sum with $\mathsf{dist}(f, g) \le \varepsilon$. By a union bound, f and g agree on all 4 points queried by the test (because each query is marginally uniformly distributed), in which case the Diamond test accepts. ◀

With completeness done, the remainder of this section is focused on showing the Diamond test's soundness as stated in Theorem 4. We restate the theorem below.

▶ **Theorem 12.** *Suppose $f : [n]^d \to \mathbb{F}_2$ passes the Diamond test with probability $1 - \varepsilon$. Then, f is $C_n\varepsilon$-close to a direct sum, for some constant C_n independent of d.*

We sketch the ideas used in the analysis. First, we will generalize to a "4-function" version of the test, which is more amenable to induction. Then, we use hypercontractivity ([1],[19], [3]) to prove Theorem 17, which establishes a dichotomy for 4-tuples of functions that pass the test with probability $1 - \varepsilon$: if (f, g, k, h) passes the test with probability ε, then either f (and g, h, k as well) is 2ε-close to a direct sum, or f is $\Omega_n(1)$-far from a direct sum. Next, we will relate the pass probability of (f, g, h, k) to the pass probability of 1-variable restrictions of f, g, k, h. This allows us to go down one dimension and apply induction. Crucially, the dichotomy theorem allows us to avoid our closeness factor deteriorating at each step of the induction, which would result in a dependence on the dimension in our result. This dichotomy based approach is inspired by the work of [7, 10], but the details of establishing the dichotomy and performing the induction are quite different.

3.1 The Four Function Diamond Test

In order to facilitate our inductive approach we define a four function version of the diamond test.

▶ **Test 13** (4-Function Diamond). Sample $x \sim \mathbb{F}_2^d$ and $a, b \sim [n]^d$. Accept iff

$$f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) + k(b) = 0.$$

The following lemma characterizes what happens when four functions pass with probability 1. It will be used later to establish the base case (in constant dimensions) of our inductive analysis.

▶ **Lemma 14.** *Suppose (f, g, h, k) passes the 4-function Diamond test with probability 1. Then f, g, k, h are direct sums.*

Proof. If (f, g, h, k) passes the 4-function Diamond test with probability 1 then for all a, x, b we have:

$$f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) + k(b) = 0 = f(b) + g(\phi_{x+\mathbb{1}}(b, a)) + h(\phi_{x+\mathbb{1}}(a, b)) + k(a).$$

Using Fact 8 to simplify we get that for all a, b,

$$f(a) + k(a) = f(b) + k(b).$$

Again using the fact that (f, g, h, k) passes the test with probability 1 we have that for all a, x, b:

$$f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) + k(b) = 0 = f(a) + g(\phi_{x+\mathbb{1}}(a, b)) + h(\phi_{x+\mathbb{1}}(b, a)) + k(b).$$

By Fact 8 this implies

$$g(\phi_x(a, b)) + g(\phi_x(b, a)) = h(\phi_x(a, b)) + h(\phi_x(b, a)).$$

Letting $c = \phi_x(a, b), d = \phi_x(b, a)$, and noting that c, d can be arbitrary (unrelated) elements of $[n]^d$, we have that for all c, d.

$$g(c) + g(d) = h(c) + h(d).$$

Combining our observations so far we conclude that there exist constants α, β such that for all $a \in [n]^d$ we have

$$f(a) = k(a) + \alpha, \quad g(a) = h(a) + \beta.$$

Thus, for all a, b we have

$$f(a) + f(b) = g(\phi_x(a, b)) + g(\phi_x(b, a)) + \alpha + \beta.$$

Taking $x = 0$ we conclude

$$f(a) + f(b) = g(a) + g(b) + \alpha + \beta.$$

Thus, there exists γ such that for all a we have

$$f(a) = g(a) + \gamma.$$

In summary, we have shown that for all a, b, x, we have

$$f(a) + f(b) = f(\phi_x(a, b)) + f(\phi_x(b, a)) + \alpha + \beta.$$

Setting $x = 0$ again shows that $\alpha + \beta = 0$. Thus, we have that f passes the (1-function) Diamond test with probability 1. By Lemma 10 this implies that f is a direct sum. Finally, recall that we have shown in the course of our proof that g, h, k all differ from f by a constant. This concludes the proof. ◀

3.2 A Dichotomy for Functions Passing the Diamond Test

We are now prepared to show Theorem 17, which states that if the 4-function Diamond test accepts (f, g, h, k) with probability exactly $1 - \varepsilon$ then f is either 2ε-close to a direct sum, or $\Omega_n(1)$-far from being a direct sum. To this end, we first establish two lemmas which will useful in the proof.

▶ **Lemma 15.** *Fix $d \in \mathbb{N}$ and functions $f, g, h : [n]^d \to \mathbb{F}_2$. Suppose f, g, h have*

$$\Pr_{a,b \sim [n]^d, x \sim \mathbb{F}_2^d}[f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) = 1] < \varepsilon.$$

Then, f, g, h are 3ε-close to constants.

Proof. Fix f, g, h as in the theorem. Performing a union bound and using Fact 8 we have

$$\Pr_{a,b,x}[f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) = f(b) + g(\phi_x(a, b)) + h(\phi_x(b, a))] > 1 - 2\varepsilon.$$

Thus, f is 2ε-close to a constant α. Then, by a union bound we have

$$\Pr_{a,b,x}[g(\phi_x(a, b)) + h(\phi_x(b, a)) = \alpha + 1] > 1 - 3\varepsilon.$$

Of course, the distribution of $(\phi_x(a, b), \phi_x(b, a))$ is the same as the distribution of (a, b). Thus, we have

$$\Pr_{a,b}[g(a) + h(b) = \alpha + 1] > 1 - 3\varepsilon.$$

By the probabilistic method this implies that g, h are each 3ε-close to constants. ◀

We will also need the following lemma, which is an extension of the classic ***small set expansion*** property of the noisy hypercube (see [24]) to *grids*.

▶ **Fact 16** ([24])**.** Fix $n \in \mathbb{N}$. There exists $\lambda_n > 0$ such that for any d, and any set $S \subseteq [n]^d$ we have

$$\Pr_{a \sim [n]^d, b \sim \mathsf{T}_{1/2}(a)}[a \in S \wedge b \in S] \leq 2\mu(S)^{1+\lambda_n}.$$

Now we are ready to establish the dichotomy theorem.

▶ **Theorem 17.** *Fix $n \in \mathbb{N}$. There exists a constant $c_n > 0$ such that the following is true for any d and any functions $f, g, h, k : [n]^d \to \mathbb{F}_2$. Suppose k has $\mathsf{dist}(k, \mathsf{directsum}) < c_n$. Then, the 4-function Diamond test rejects (f, g, h, k) with probability at least $\mathsf{dist}(k, \mathsf{directsum})/2$.*

Proof. Fix f, g, k, h as described in the theorem statement. Let χ be the closest direct sum to k. Then,

$$\chi(a) + \chi(b) = \chi(\phi_x(a, b)) + \chi(\phi_x(b, a)).$$

Thus, our goal is equivalent to showing that

$$\Pr_{a,x,b}[(\chi + f)(a) + (\chi + g)(\phi_x(a, b)) + (\chi + h)(\phi_x(b, a)) \neq (\chi + k)(b)] \geq \mathsf{dist}(k, \mathsf{directsum})/2.$$

Thus, it suffices to consider the case that the closest direct sum to k is the zero function. We assume this for the remainder of the proof. Then, $\mathsf{dist}(k, \mathsf{directsum}) = \mu(k)$. We are already done unless

$$\Pr_{a,x,b}[f(a) + g(\phi_x(a, b)) + h(\phi_x(b, a)) = 1] \leq 2\mu(k). \tag{2}$$

Thus, we may assume that Equation (2) is true. Now, by Lemma 15 we have that f, g, h are $6\mu(k)$-close to constants. Without loss of generality we may assume g, h are both closer to 0 than to 1; otherwise we can add 1 to both g, h, which does not affect the sum $g(\phi_x(a,b)) + h(\phi_x(b,a))$. If f were closer to 1 than to 0 then we would have

$$\Pr_{(a,x,b)}[f(a) = 1, g(\phi_x(a,b)) = h(\phi_x(b,a)) = 0] \geq 1 - 18\mu(k) > 2\mu(k),$$

with the final inequality holding by our assumption that $\mu(k) < c_n$ (we will choose $c_n < .001$); but this would contradict Equation (2), so it cannot happen. Thus, f, g, h are all $6\mu(k)$-close to 0.

Now, we return to analyzing the probability that the 4-function Diamond test rejects (f, g, h, k); we write $\Pr[\mathsf{rej}]$ to denote the probability. Then,

$$\begin{aligned}
\Pr[\mathsf{rej}] &\geq \Pr_{a,x,b}[k(b) = 1 \wedge f(a) = g(\phi_x(a,b)) = h(\phi_x(b,a)) = 0] \\
&\geq \mu(k) - \Pr_{a,b}[k(b) = f(a) = 1] - \Pr_{a,x,b}[k(b) = g(\phi_x(a,b)) = 1] - \Pr_{a,x,b}[k(b) = h(\phi_x(b,a)) = 1] \\
&\geq \mu(k) - \mu(k)\mu(f) - \Pr_{a,b\sim\mathsf{T}_{1/2}(a)}[k(b) = g(a) = 1] - \Pr_{a,b\sim\mathsf{T}_{1/2}(a)}[k(b) = h(a) = 1].
\end{aligned}$$

Then, using Fact 16 we have

$$\Pr[\mathsf{rej}] \geq \mu(k)(1 - \mu(f)) - (\mu(k) + \mu(g))^{1+\lambda_n} - (\mu(k) + \mu(h))^{1+\lambda_n}),$$

where $\lambda_n > 0$ is the constant from Fact 16. Recalling that $\mu(f), \mu(g), \mu(h) \leq 6\mu(k)$ we have:

$$\Pr[\mathsf{rej}] \geq \mu(k)(1 - \mu(k) - 100\mu(k)^{\lambda_n}).$$

Thus, there is some constant c_n (dependent only on n) such that if $\mu(k) \leq c_n$ then the rejection probability in is at least $\mu(k)/2$, as desired. ◀

3.3 A Lemma Concerning Restrictions

To complete our analysis of the Diamond test, we will relate the Diamond test to lower dimensional Diamond tests, and argue inductively. The Dichotomy lemma just established will be the key to ensuring that our soundness parameter does not deteriorate as we perform the induction. First, we need a lemma that controls the restrictions of a function which is far from a direct sum; namely, we show that a function which is far from a direct sum cannot have too many restrictions which are close to direct sums.

For some set $R \subseteq [n]^d$, we let $f|_R$ denote the restriction of f to R, so that $f|_R : S \to \mathbb{F}_2$ is given by $f|_R(x) = f(x)$ for all $x \in R$. We will specifically be considering restrictions where R is of the form $R = \{x \in [n]^d \mid x_i = j\}$ for some $1 \leq i \leq d$ and some $j \in [n]$. We call such R, *one-variable restrictions*. Throughout this subsection we use $\mathbb{1}$ to denote the indicator of an event. The goal of this subsection is to establish the following Lemma.

▶ **Lemma 18.** *Suppose there are $100n^2/\varepsilon$ one-variable restrictions R such that $f|_R$ is ε-close to a direct sum on R. Then, f is $O(\varepsilon)$-close to a direct sum.*

We will need Dinur and Golubev's result regarding the soundness of the square in the cube test, which we state below.

▶ **Theorem 19.** *[13] Let $f : [n]^d \to \mathbb{F}_2$ satisfy,*

$$\Pr_{a,b\in[n]^d, x,y,z\in\{0,1\}^d}[f(a) + f(\phi_x(a,b)) + f(\phi_y(a,b)) + f(\phi_{x+y}(a,b)) = 0] \geq 1 - \varepsilon.$$

Then f is $O(\varepsilon)$-close to a direct sum.

At a high level our proof of Lemma 18 follows the strategy of the plane versus point analysis of [25]. Suppose that there are many restrictions R_i with local direct sums $f_i : R_i \to \mathbb{F}_2$ such that f_i and f are close. We will first show that many — in fact at least a constant fraction — of these f_i's are actually consistent with each other. Using these consistent f_i's we can then define a function F on the whole space and argue that 1) F is close to f and 2) F passes the square in the cube test with high probability and is thus close to a direct sum. For both 1) and 2) we are crucially using the fact that F is defined using many consistent direct sums, so for nearly every point x in the domain, there are in fact many of these local direct sums defined on x.

3.3.1 A Sampling Lemma

In order to carry out the mentioned strategy, we must first show a sampling lemma. This lemma roughly states that given a large enough set of restrictions R_i, the measure on $[n]^d$ produced by first choosing a random R_i and then choosing a random $x \in R_i$ is nearly equal to the uniform measure up to constant factors (which will be negligible).

Let $\mathcal{R}$ be a set of one variable restrictions and let $M = |\mathcal{R}|$. Define the measure $\nu_{\mathcal{R}}$ on points $[n]^d$ via the following sampling procedure:

- Choose $R \in \mathcal{R}$.
- Choose $x \in R$.

Letting $N_{\mathcal{R}}(x) = |\{R \in \mathcal{R} \mid R \ni x\}|$, it is clear that,

$$\nu_{\mathcal{R}}(x) = \frac{N_{\mathcal{R}}(x)}{M} \cdot \frac{1}{n^{d-1}}.$$

For a subset of points $S \subseteq [n]^d$, we define

$$\nu_{\mathcal{R}}(S) = \sum_{x \in S} \nu_{\mathcal{R}}(x).$$

We also let μ denote the standard uniform measure on $[n]^d$, so for a set of points $S \subseteq [n]^d$, we have $\mu(S) = \frac{|S|}{n^d}$. At times we will also use μ to denote the uniform measure over grids of dimension other than d. It will be clear from context when this is the case and we remark that one should think of the dimension in this subsection as being arbitrary.

▶ **Lemma 20.** *Let $\mathcal{R}$ be a set of one-variable restrictions of size M. Then,*

$$\mathbb{E}_x[N_{\mathcal{R}}(x)] = \frac{M}{n} \quad \textit{and} \quad \mathsf{Var}(N_{\mathcal{R}}(x)) \leq \frac{M}{n}.$$

Proof. For the first part, we have, by linearity of expectation:

$$\mathbb{E}_x[N_{\mathcal{R}}] = \sum_{R \in \mathcal{R}} \mathbb{E}_x[\mathbb{1}(x \in R)] = \frac{M}{n}.$$

Towards the second part, we write

$$\begin{aligned}
\mathbb{E}[N_{\mathcal{R}}^2] &= \sum_{R_1, R_2 \in \mathcal{R}} \mathbb{E}_x[\mathbb{1}(x \in R_1) \cdot \mathbb{1}(x \in R_2)] \\
&= \sum_{R \in \mathcal{R}} \mathbb{E}_x[\mathbb{1}(x \in R)] + \sum_{R_1 \neq R_2} \mathbb{E}_x[\mathbb{1}(x \in R_1) \cdot \mathbb{1}(x \in R_2)] \\
&\leq \frac{M}{n} + \frac{M^2}{n^2}.
\end{aligned}$$

It follows that,

$$\mathrm{Var}(N_{\mathcal{R}}) = (\mathbb{E}_x[N_{\mathcal{R}}])^2 - \mathbb{E}[N_{\mathcal{R}}^2] \leq \frac{M}{n}.$$

◀

Applying Chebyshev's inequality, we get the following.

▶ **Lemma 21.** *For any $c > 0$, it holds that*

$$\Pr_x\left[\left|N_{\mathcal{R}}(x) - \frac{M}{n}\right| \geq c \cdot \frac{M}{n}\right] \leq \frac{n}{c^2 \cdot M}.$$

Proof. This lemma follows from a direct application of Chebyshev's inequality combined with the variance bound in Lemma 20. ◀

▶ **Lemma 22.** *For any set of points $S \subseteq [n]^d$ and any set $\mathcal{R}$ of one-variable restrictions of size M, we have*

$$\frac{1}{2}\left(\mu(S) - \frac{4n}{M}\right) \leq \nu_{\mathcal{R}}(S) \leq 2\mu(S) + \frac{5n}{M}.$$

Proof. For each integer i, let

$$m_i = \left|\left\{x \in [n]^d \mid 2^i\frac{M}{n} \leq N_{\mathcal{R}}(x) < 2^{i+1}\frac{M}{n}\right\}\right|.$$

By Lemma 20 we get that

$$\frac{m_i}{n^d} \leq \frac{1}{(2^i-1)^2 M/n}.$$

Towards the upper bound, we expand $\nu_{\mathcal{R}}(S)$ and perform a dyadic partitioning:

$$\nu_{\mathcal{R}}(S) = \sum_{x\in S}\nu_{\mathcal{R}}(x) = \sum_{x\in S}\frac{N_{\mathcal{R}}(x)}{Mn^{d-1}} \leq \frac{1}{Mn^{d-1}}\left(2\cdot\frac{M}{n}\cdot|S| + \sum_{i=1}^{\infty} m_i 2^{i+1}\cdot\frac{M}{n}\right).$$

The first term in the parenthesis contributes $2\mu(\mathcal{S})$, whereas the contribution of the second summand can be upper bounded by

$$\frac{1}{n^d}\cdot\sum_{i=1}^{\infty} m_i 2^{i+1} \leq \frac{n}{M}\cdot\sum_{i=1}^{\infty}\frac{2^{i+1}}{(2^i-1)^2} \leq \frac{5n}{M}$$

For the lower bound, we have

$$\nu_{\mathcal{R}}(S) = \sum_{x\in S}\nu_{\mathcal{R}}(x) \geq \sum_{x\in S, N_{\mathcal{R}}\geq M/(2n)}\nu_{\mathcal{R}}(x) \geq |\{x\in S \mid N_{\mathcal{R}}(x) \geq M/(2n)\}|\cdot\frac{1}{2n^d}.$$

To lower bound the last term above, we use Lemma 21 to get

$$\Pr_x\left[N_{\mathcal{R}}(x) < \frac{M}{2n}\right] \leq \frac{4n}{M}.$$

It follows that

$$|\{x\in S \mid N_{\mathcal{R}}(x) \geq M/(2n)\}| \geq n^d\left((\mu(S) - \frac{4n}{M}\right).$$

Putting everything together, we get the desired lower bound:

$$\nu_{\mathcal{R}}(S) \geq \frac{1}{2}\left(\mu(S) - \frac{4n}{M}\right).$$

◀

3.3.2 Proof Lemma 18

Suppose $\mathcal{R} = \{R_1, \dots, R_M\}$ is the set of one-variable restrictions on which f is ε-close to a direct sum. For each R_i, let f_i be this direct sum, so

$$\mathsf{dist}(f_i, f|_{R_i}) \le \varepsilon,$$

for each $1 \le i \le M$. By assumption $M \ge 100n^2/\varepsilon$. Note that if ε is large — say $\varepsilon \ge \frac{1}{100}$ — then the result is trivial, so for the remainder of the proof we suppose $\varepsilon < \frac{1}{100}$.

As a useful fact, we show that the set of direct sum functions has minimum distance $\frac{1}{n}$.

▶ **Lemma 23.** *For any two distinct direct sums f and g, we have*

$$\mathsf{dist}(f, g) \ge \frac{1}{n}.$$

Proof. Write $f(x) = \sum_{i=1}^d L_i(x_i)$ and $g(x) = \sum_{i=1}^d L'_i(x_i)$. First suppose for each i, $L_i - L'_i$ is constant, then since $f(x) \neq g(x)$, we have $\delta(f, g) = 1$ and we are done.

Suppose that there is some i such that $L_i - L'_i$ is not the constant function and without loss of generality, let this i be 1. Then,

$$\Pr_{x \in [n]^d}[f(x) \neq g(x)] = \mathbb{E}_{x_2, \dots, x_d}\left[\Pr_{x_1}[L_1(x) - L'_1(x) \neq \sum_{i=2}^d L_i(x_i) - L'_i(x_i)]\right].$$

To conclude, note that for any $x_2, \dots, x_d$, the inner probability on the right hand side above is at least $1/n$ since $L_1 - L'_1$ is not the constant function. ◀

We now construct the following graph $G = (V, E)$. The vertex set is $V = \{1, \dots, M\}$ while the set of edges E consists of all pairs (i, j) such that the functions f_i and f_j are consistent:

$$f_i|_{R_i \cap R_j} = f_j|_{R_i \cap R_j}.$$

If $R_i \cap R_j$ is empty, then we also consider f_i and f_j consistent and have $(i, j) \in E$. We will first show that G contains a large clique. Call a graph is *transitive* if it is an edge disjoint union of cliques. Also define

$$\beta(G) = \max_{(i,j) \notin E} \Pr_{k \in V}[(i, k), (j, k) \in E].$$

Note that G is transitive if and only if $\beta(G) = 0$. The following lemma from [25] gives a sort of approximate version of this fact and states that if $\beta(G)$ is small then G is almost a transitive graph.

▶ **Lemma 24.** *The graph G can be made transitive by removing at most $3\sqrt{\beta(G)}|V|^2$ edges.*

In our next two lemmas we show that G has many edges, and then bound $\beta(G)$. This will establish that G contains a large clique, which corresponds to many local direct sums f_i that are consistent with one another.

▶ **Lemma 25.** *The graph G has at least $0.9M^2$ edges.*

Proof. For each R_i, let $S_i \subseteq R_i$ be the set of points on which f_i and f differ. We have that

$$\frac{|S_i|}{|R_i|} \le \varepsilon,$$

by assumption.

Now fix an R_i. We will apply Lemma 22 inside of R_i, which is isomorphic to the grid $[n]^{d-1}$. For each j, let $R'_j = R_i \cap R_j$. Let $\mathcal{R}' = \{R'_j \mid \frac{R'_j \cap S_i}{R'_j} \geq 3\varepsilon\}$, and let $M' = |\mathcal{R}'|$. Now consider the size of S_i inside of R_i under the measure $\nu_{\mathcal{R}'}$. We have,

$$\nu_{\mathcal{R}'}(S_i) \geq 3\varepsilon.$$

By Lemma 22

$$3\varepsilon \leq \nu_{R'}(S_i) \leq 2\varepsilon + 5n/M'.$$

It follows that,

$$M' \leq \frac{5n}{\varepsilon}.$$

Thus,

$$\Pr_{R_i, R_j}\left[\frac{|R_i \cap R_j \cap S_i|}{|R_i \cap R_j|} \geq 3\varepsilon\right] \leq \frac{5n}{\varepsilon M}.$$

By a union bound, it follows that with probability at least $1 - \frac{10n}{\varepsilon M}$, we have that both

$$\frac{|R_j \cap R_i \cap S_i|}{|R_j \cap R_i|} \leq 3\varepsilon \quad \text{and} \quad \frac{|R_j \cap R_i \cap S_j|}{|R_j \cap R_i|} \leq 3\varepsilon.$$

For such i, j, we get that

$$\Pr_{x \in R_i \cap R_j}[f_i(x) = f_j(x)] \geq 1 - \Pr_{x \in R_i \cap R_j}[x \in S_i \cup S_j] \geq 1 - 6\varepsilon > \frac{1}{n},$$

and $f_i|_{R_i \cap R_j} = f_j|_{R_i \cap R_j}$ by Lemma 23. Thus choosing i, j randomly, we get that they are adjacent in G with probability at least $1 - \frac{10n}{\varepsilon M} \geq 0.9$. ◀

▶ **Lemma 26.** *Suppose R_i, R_j are distinct one-variable restrictions such that $f_i|_{R_i \cap R_j} \neq f_j|_{R_i \cap R_j}$. Then,*

$$\Pr_{R_k}[f_i|_{R_i \cap R_j \cap R_k} = f_j|_{R_i \cap R_j \cap R_k}] \leq \frac{5n^2}{M} \leq \frac{\varepsilon}{20}.$$

Proof. Let $W = R_i \cap R_j$. For each k let $R'_k = R_k \cap W$, let $f'_i = f_i|_W$, and let $f'_j = f_j|_W$.

Throughout the proof we will consider W as our ambient space and we will apply Lemma 22 inside of W. Note that we may do so because W is isomorphic to the grid $[n]^{d-2}$. Define the following set of one-variable restrictions inside W:

$$\mathcal{R}' = \{W \cap R_k \neq \emptyset \mid R_k \in \mathcal{R}, f'_i|_{R'_k} = f'_j|_{R'_k}\}.$$

We will show that $|\mathcal{R}'|$ cannot be too large. Indeed, let M' denote its size and let $S \subseteq W$ be

$$S = \{x \in W \mid f'_i(x) \neq f'_j(x)\}.$$

We have that $\frac{|S|}{|W|} \geq \frac{1}{n}$. On the other hand, by definition of $\mathcal{R}'$, we have that $\nu_{R'}(S) = 0$. Therefore by Lemma 22:

$$0 \geq \frac{1}{2}\left(\frac{1}{n} - \frac{4n}{M'}\right).$$

It follows that

$$M' \leq 4n^2.$$

Finally there can be up to $2n$ additional restrictions R_k such that $R_k \cap R_i = \emptyset$ or $R_k \cap R_j = \emptyset$. The result follows. ◀

Combining Lemma 25 and Lemma 26, we get that G contains a transitive subgraph with at least $0.8M^2$ edges. Let $\mathcal{C}_1, \ldots, \mathcal{C}_J$ be the edge disjoint union cliques of this transitive subgraph and say $\mathcal{C}_1$ is the largest one. It follows that,

$$0.8M^2 \leq \sum_{I=1}^{J} |\mathcal{C}_i|^2 \leq |\mathcal{C}_1| \cdot M,$$

and G contains a clique of size $M_1 = 0.8M$. From now we let $\mathcal{C}$ denote this clique and let $\mathcal{C} = \{1, \ldots, M_1\}$. We define the following function F using the functions $f_1, \ldots, f_{M_1}$: set $F(x) = f_j(x)$ if there is some $j \in \mathcal{C}_1$ such that $x \in R_j$, and $F(x) = 0$ otherwise.

To conclude, we will show that F is close to f and F is close to a direct sum. This will establish that f is close to a direct sum.

▶ **Lemma 27.** *F is 3ε-close to f.*

Proof. Let $S = \{x \in [n]^d \mid f(x) \neq F(x)\}$. Since f is ε-close to F on the restrictions R_i for all $i \in \mathcal{C}$, we have

$$\nu_{\mathcal{C}}(S) \leq \varepsilon.$$

By Lemma 22, it follows that,

$$\varepsilon \geq \frac{1}{2}\left(\mu(S) - \frac{4n}{0.8M}\right),$$

and

$$\mu(S) \leq 3\varepsilon.$$

◀

▶ **Lemma 28.** *F is $O(\varepsilon)$-close to a direct sum.*

Proof. We will show that F passes the square in a cube test with high probability. Recall this test operates by choosing $a, b \in [n]^d$, $x, y \in \{0,1\}^d$ and checking if F satisfies,

$$F(a) + F(\phi_x(a,b)) + F(\phi_y(a,b)) + F(\phi_{x+y}(a,b)) = 0.$$

Note that F passes if there is some $i \in \mathcal{C}$ such that $a, b \in R_i$. Indeed, in this case all of the queried points are in R_i, so F and the direct sum f_i agree on all of the queried points and

$$F(a)+F(\phi_x(a,b))+F(\phi_y(a,b))+F(\phi_{x+y}(a,b)) = f_i(a)+f_i(\phi_x(a,b))+f_i(\phi_y(a,b))+f_i(\phi_{x+y}(a,b)) = 0.$$

To bound the probability that $i \in \mathcal{C}$ such that $a, b \in R_i$, let $N'(x) = |\{i \in \mathcal{C} \mid x \in R_i\}|$. By Lemma 21,

$$\Pr_a\left[N'(a) \geq \frac{M}{1.6n}\right] \leq \frac{5n}{M}.$$

Now condition on a satisfying $N'(a) \geq \frac{M}{1.6n}$ and let $\mathcal{C}'' = \{i \in \mathcal{C} \mid a \in R_i\}$. We label $\mathcal{C}'' = \{1, \ldots, M''\}$ where $M'' \geq \frac{M}{1.6n}$ and let $N''(x) = |\{i \in \mathcal{C}'' \mid x \in R_i\}|$. Then by Lemma 21 again,

$$\Pr_b[N''(b) \geq 0] \leq \frac{n}{M''} \leq \frac{1.6n^2}{M}.$$

By a union bound, with probability at least $1 - \frac{7n^2}{M}$ we have that a and b are both contained in some R_i with $i \in \mathcal{C}$. It follows that F passes the square in a cube test with probability at least $1 - \frac{7n^2}{M}$. By Theorem 19, F is $O(\frac{n^2}{M}) = O(\varepsilon)$-close to a direct sum. ◀

Proof of Lemma 18. The result follows by combining Lemma 27, Lemma 28 and using the triangle inequality. ◀

3.4 Soundness of the Diamond Test

Equipped with Theorem 17 we complete the proof of Theorem 4. Ultimately, we wish to conclude that if the 4-function Diamond test rejects (f, g, h, k) with probability exactly ε, then f, g, h, k are all $O(\varepsilon)$-close to direct sums. However, what we have thus far only allows us to to conclude that f, g, k, h are each *either* c_n-far from being a direct sum, *or* 2ε-close to a direct sum. Our next theorem rules out the former possibility.

▶ **Theorem 29.** *Fix $n \in \mathbb{N}$. Let c_n be the constant from Theorem 17. There is a constant δ_n such that the following holds. Fix $d \in \mathbb{N}$, and functions $f, g, h, k : [n]^d \to \mathbb{F}_2$. Suppose f is c_n-far from being a direct sum. Then, the 4-function Diamond test rejects (f, g, h, k) with probability at least δ_n.*

Proof. By Lemma 18 there exists a constant K_n such that if f has at least $100n^2/c_n$ restrictions which are c_n/K_n-close to affine, then f is c_n-close to affine. Define $N_n = \lceil 100n^2/c_n \rceil + 1$ and $\delta_n = \min(c_n/(2K_n), n^{-3N_n})$. We will prove the theorem by induction on the dimension d.

The base case is $d \leq N_n$. If $d \leq N_n$ then by Lemma 10 the 4-function Diamond test rejects (f, g, h, k) with non-zero probability since f is not a direct sum. But, any non-zero probability over events determined by a, b, x is at least $1/n^{3N_n} \geq \delta_n$.

Now, fix $d > N_n$, and assume the theorem for functions on $\mathbb{F}_2^{d-1}$. Let $f\,|_{i,\sigma}\colon [n]^{d-1} \to \mathbb{F}_2$ denote the function obtained by taking f and fixing $x_i = \sigma$. Then, there is some $i \in [d]$ such for all $\sigma \in [n]$ the restriction $f\,|_{i,\sigma}$ is c_n/K_n-far from a direct sum. To analyze the probability that the 4-function Diamond test rejects (f, g, h, k), we consider a partition of the probability space into $2n^2$ disjoint events. The events are the value in $[n]^2 \times \mathbb{F}_2$ that our sampled (a_i, b_i, x_i) takes. Now, consider the probability that the 4-function Diamond test rejects (f, g, h, k), conditional on $(a_i, b_i, x_i) = (\alpha, \beta, \xi)$ for some α, β, ξ. Conditional on this event, we will sample $a, b \in [n]^{d-1}, x \in \mathbb{F}_2^{d-1}$ and check:

$$f\,|_{i,\alpha}\,(a) + g\,|_{i,\phi_\xi(\alpha,\beta)}\,(\phi_x(a, b)) = h\,|_{i,\phi_\xi(\beta,\alpha)}\,(\phi_x(b, a)) + k\,|_{i,\beta}\,(b).$$

This is precisely a 4-function Diamond test on $(d-1)$-dimensional functions. Now we consider two cases.
Case 1: $f\,|_{i,\alpha}$ is c_n-far from being a direct sum.
Then by the inductive hypothesis this $(d-1)$-dimension test will reject with probability at least δ_n.
Case 2: $\mathsf{dist}(f\,|_{i,\alpha}, \mathsf{directsum}) \in [c_n/K_n, c_n]$.

Then, by Theorem 17 this $(d-1)$-dimensional test rejects with probability at least $c_n/(2K_n) \geq \delta_n$.

We have seen that in all cases, the lower dimensional 4-function Diamond tests reject with probability at least δ_n. Thus, the actual 4-function Diamond test rejects with probability δ_n as well. ◀

Combining Theorem 17 and Theorem 29 we have get the following corollary and conclude the proof of Theorem 4.

▶ **Corollary 30.** *If (f, g, h, k) passes the 4-function Diamond test with probability $1 - \varepsilon$, then f, g, h, k are each $O_n(\varepsilon)$-close to direct sums.*

Proof. The result follows immediately from Theorem 17 and Theorem 29. ◀

We can now conclude the proof of Theorem 4 and establish soundness for the Diamond test.

Proof of Theorem 4. The result follows from Corollary 30 by setting $f = g = h = k$. ◀

3.5 The Fourier Analytic Interpretation of the Diamond Test

In this section, we give a Fourier analytic interpretation of the Diamond test. Let Even denote the class of functions $f : \mathbb{F}_2^d \to \mathbb{F}_2$ satisfying $f(x + \mathbb{1}) = f(x)$ and let Odd denote the class of functions $f : \mathbb{F}_2^d \to \mathbb{F}_2$ satisfying $f(x + \mathbb{1}) = f(x) + 1$. Let $\mathsf{EvenOrOdd} = \mathsf{Even} \cup \mathsf{Odd}$. We note that $\mathsf{EvenOrOdd}$ has a nice Fourier analytic interpretation: $f \in \mathsf{EvenOrOdd}$ implies that $(-1)^f$ has either all its Fourier-mass on even-sized characters, or all of its Fourier-mass on odd-sized characters. Given a function $f : \mathbb{F}_2^d \to \mathbb{F}_2$, define random variable $\mathcal{R}(f)$ to be a random restriction of f obtained as follows: for each $i \in [d]$ independently, with probability $1/4$ restrict x_i to 0, with probability $1/4$ restrict x_i to 1, and with probability $1/2$ do not restrict x_i. We now show that the Diamond test's correctness is equivalent to the following Fourier analytic fact

▶ **Theorem 31.**

$$\mathsf{dist}(f, \mathsf{affine}) \leq O(\mathbb{E}[\mathsf{dist}(\mathcal{R}(f), \mathsf{EvenOrOdd})])$$

In fact, we show something stronger: there is a very tight quantitative relationship between

$$\mathbb{E}[\mathsf{dist}(\mathcal{R}(f), \mathsf{EvenOrOdd})]$$

and the probability that the Diamond test rejects f.

▶ **Lemma 32.** *Fix f. Let ε be the probability that the Diamond test rejects f, and let*

$$\delta = \mathbb{E}[\mathsf{dist}(\mathcal{R}(f), \mathsf{EvenOrOdd})].$$

Then,

$$2\delta \leq \varepsilon \leq 4\delta.$$

Proof. Fix f, ε, δ as in the lemma statement. Given a, b define $C_{a,b} = \left\{\phi_x(a, b) \mid x \in \mathbb{F}_2^d\right\}$ and $f_{a,b}(x) = f(\phi_x(a, b))$. We have:

$$\varepsilon = \Pr_{a,b,x}[f_{a,b}(x) + f_{a,b}(x + \mathbb{1}) \neq f_{a,b}(0) + f_{a,b}(\mathbb{1})]. \tag{3}$$

▷ Claim 33. $\varepsilon \leq 4\delta$.

Let $\delta_{a,b} = \mathsf{dist}(f_{a,b}, \mathsf{EvenOrOdd})$. By a union bound we have

$$\Pr_{z,x}[f_{a,b}(x+\mathbb{1}) + f_{a,b}(x) \neq f_{a,b}(z+\mathbb{1}) + f_{a,b}(z)] \leq 4\delta_{a,b}.$$

By averaging, $\delta = \mathbb{E}_{a,b}[\delta_{a,b}]$. Thus,

$$\Pr_{a,b,z,x}[f_{a,b}(x+\mathbb{1}) + f_{a,b}(x) \neq f_{a,b}(z+\mathbb{1}) + f_{a,b}(z)] \leq 4\delta. \tag{4}$$

Now, observe that for random a, b, z, x the following two distributions are the same:
1. $(\phi_x(a,b), \phi_x(b,a), a, b)$
2. $(\phi_x(a,b), \phi_x(b,a), \phi_z(a,b), \phi_z(b,a))$.

Thus Equation (4) implies $\varepsilon \leq 4\delta$.

▷ Claim 34. $\varepsilon \geq 2\delta$.

For any function g,

$$\Pr_x[g(x) \neq g(x+\mathbb{1})] = 2\mathsf{dist}(g, \mathsf{Even}),$$
$$\Pr_x[g(x) = g(x+\mathbb{1})] = 2\mathsf{dist}(g, \mathsf{Odd}).$$

Thus, for any a, b we have

$$\Pr_x[f_{a,b}(x) + f_{a,b}(x+\mathbb{1}) \neq f_{a,b}(0) + f_{a,b}(\mathbb{1})] \geq 2\mathsf{dist}(f_{a,b}, \mathsf{EvenOrOdd}).$$

Averaging over a, b on both sides gives $\varepsilon \geq 2\delta$.

◀

Proof of Theorem 31. The result follows from the soundness of the Diamond Test, Theorem 4 ◀

4 The Affine on Subcube Test

In this section we consider a class of direct sum tests that generalizes the Square in Cube test of [13]. We call these tests ***Affine on Subcube*** tests. Indeed, the Square in Cube test can be viewed as performing the BLR affinity test inside of a random induced hypercube inside of $[n]^d$. We show that in fact testing for affinity on a random subcube with *any* affinity tester also works as a direct sum test.

▶ **Test 35** (Affine on Subcube)**.** Sample $a, b \in [n]^d$ and run an affinity test that has soundness function $s(\delta) = \Omega(\delta)$ on the function $x \mapsto f(\phi_x(a,b))$.

The flexibility of being able to use any affinity tester also enables us to obtain a direct sum test with additional properties such as erasure resiliency, which we discuss further in Section 4.2, and less use of randomness, which we discuss in Section 4.1. Furthermore, we believe that our analysis for going from local direct sums to a global direct sum is simpler than that of [13] and thus noteworthy on its own.

One final benefit of our analysis is that it fixes an error in [13]'s analysis. Specifically, they establish that a function f passing the Square in Cube test with probability $1-\varepsilon$ is *locally* close to a direct sum (this is basically Lemma 39 in our presentation), but they do not show how to combine these local direct sums into a global direct sum.

We now show that the Affine on Subcube test is a valid direct sum tester. Once again, completeness is clear. If f is a direct sum, then the Affine on Subcube test passes with probability 1. Our focus is then on showing soundness.

▶ **Theorem 36.** *Fix $f : [n]^d \to \mathbb{F}_2, \varepsilon \geq 0$ and an affinity test T. If f passes the Affine on Subcube test (instantiated with T) with probability $1 - \varepsilon$, then f is $C\varepsilon$-close to a direct sum, for some constant C independent of n, d.*

Proof. Fix f as in the theorem statement. That is, satisfying

$$\mathbb{E}_{a,b\sim[n]^d}[\mathsf{dist}(x \mapsto f(\phi_x(a,b)), \mathsf{affine})] \leq O(\varepsilon). \tag{5}$$

We re-interpret Equation (5) in three steps. First, observe that Equation (5) is equivalent to

$$\mathbb{E}_{a,b\sim[n]^d, z\sim\mathbb{F}_2^d}[\mathsf{dist}(x \mapsto f(\phi_z(a,b)) + f(\phi_{x+z}(a,b)), \mathsf{linear})] \leq O(\varepsilon). \tag{6}$$

Next, observe that we can rewrite $\phi_{x+z}(a,b)$ as follows:

$$\phi_{x+z}(a,b) = \phi_x(\phi_z(a,b), \phi_z(b,a)).$$

In light of this observation, Equation (6) is equivalent to

$$\mathbb{E}_{a,b\sim[n]^d, z\sim\mathbb{F}_2^d}[\mathsf{dist}(x \mapsto f(\phi_z(a,b)) + f(\phi_x(\phi_z(a,b), \phi_z(b,a))), \mathsf{linear})] \leq O(\varepsilon). \tag{7}$$

Of course, the distribution of $(\phi_z(a,b), \phi_z(b,a))$ is the same as the distribution of (a,b). So, we can rewrite Equation (7) as:

$$\mathbb{E}_{a,b\sim[n]^d}[\mathsf{dist}(x \mapsto f(a) + f(\phi_x(a,b)), \mathsf{linear})] \leq O(\varepsilon). \tag{8}$$

This interpretation of the test is more convenient to work with.

For each a, b define f_{ab} to be the function $x \mapsto f(a) + f(\phi_x(a,b))$, and let ε_{ab} be the probability that f_{ab} fails the linearity test. Clearly $\mathbb{E}[\varepsilon_{ab}] = \mathbb{E}[\varepsilon_a] \leq O(\varepsilon)$. Because f_{ab} passes the linearity test with probability $1 - \varepsilon_{ab}$, there is vector $F^a(b) \in \mathbb{F}_2^d$ such that

$$\Pr_x[f_{ab}(x) = \langle F^a(b), x\rangle] > 1 - O(\varepsilon_{ab}), \tag{9}$$

where $\langle v, w\rangle = \sum_i v_i w_i$. We will now show that $b \mapsto F^a(b)$ is close to a direct product, by proving that this function passes a certain direct product test. Dinur and Golubev [13] give the following direct product test (which is slightly more convenient here than the original direct product test of [16]):

▶ **Test 37** (Direct Product Test)**.** Sample $x \in [n]^d$. Sample a set A by including each $i \in [d]$ in A with probability $3/4$. Sample y by setting $y_i = x_i$ if $i \in A$, and otherwise sampling y_i uniformly from $[n_i]$. Accept iff $f(x)_i = f(y)_i$ for all $i \in A$.

Using [11], Dinur and Golubev show

▶ **Fact 38.** If f passes Test 37 with probability $1-\varepsilon$, then f is $O(\varepsilon)$-close to a direct product.

Now, we use Fact 38 to show that $b \mapsto F^a(b)$ is close to a direct product.

▶ **Lemma 39.** *$b \mapsto F^a(b)$ is $O(\varepsilon_a)$-close to a direct product.*

Proof. Given b, b' define $\mathcal{D}_{bb'}$ on $\mathbb{F}_2^d$ as follows. If $x \sim \mathcal{D}_{bb'}$ then for each $i \in [d]$ independently, x_i is determined as follows:

If $b_i \neq b'_i$, then $x_i = 0$. If $b_i = b'_i$, then $x_i = 1$ with probability $2/3$ and 0 otherwise.

Define $\mathsf{eq}_{3/4} = \mathsf{T}_{(3n/4-1)/(n-1)}$. If $b' \sim \mathsf{eq}_{3/4}(b)$, then for each i independently, $b_i = b'_i$ with probability $3/4$. Now, observe that if we sample $b' \sim \mathsf{eq}_{3/4}(b)$ and $x \sim \mathcal{D}_{bb'}$, then the values x_i are independent uniformly random bits. However, if we choose b', x in this manner, then the joint distribution of (b, b', x) satisfies

$$\phi_x(a, b) = \phi_x(a, b').$$

This is because we select a_i whenever $b_i \neq b'_i$. Then, performing a union bound on Equation (9) (and averaging away the dependence on b) we have that

$$\Pr_{b\sim[n]^d, b'\sim\mathsf{eq}_{3/4}(b), x\sim\mathcal{D}_{bb'}}[\langle F^a(b), x\rangle = f_{ab}(x) = f_{ab'}(x) = \langle F^a(b'), x\rangle] > 1 - O(\varepsilon_a). \tag{10}$$

For each a, b, b', define

$$\delta_{abb'} = \Pr_{x\sim\mathcal{D}_{bb'}}[\langle F^a(b) + F^a(b'), x\rangle \neq 0].$$

By Equation (10), $\mathbb{E}_{b\sim[n]^d, b'\sim\mathsf{eq}_{3/4}(b)}[\delta_{abb'}] < O(\varepsilon_a)$. Intuitively, if $\delta_{abb'}$ is small then $F^a(b), F^a(b')$ must have some "agreement". More precisely we have:

▷ Claim 40. If $\delta_{abb'} < 1/3$, then for all i where $b_i = b'_i$ we have $F^a(b)_i = F^a(b')_i$

Proof. Suppose to the contrary that there is some $i \in [d]$ with $b_i = b'_i$ but $F^a(b)_i \neq F^a(b')_i$. If $x \sim \mathcal{D}_{bb'}$ then there is a $1/3$ chance of $x_i = 0$ and a $2/3$ chance of $x_i = 1$, and this is independent of the values of x_j for $j \neq i$. This means that $\langle F^a(b) + F^a(b'), x\rangle$ takes on either value in $\mathbb{F}_2$ with probability at least $1/3$, implying that $\delta_{abb'} \geq 1/3$. ◀

Now, by Markov's Inequality we have

$$\Pr_{b\sim[n]^d, b'\sim\mathsf{eq}_{3/4}(b)}[\delta_{abb'} \geq 1/3] \leq 3\mathbb{E}_{b,b'}[\delta_{abb'}] < O(\varepsilon_a). \tag{11}$$

Given b, b' let $\overline{\Delta}(b, b')$ denote the set of coordinates i where $b_i = b'_i$. By Equation (11) we have

$$\Pr_{b\sim[n]^d, b'\sim\mathsf{eq}_{3/4}(b)}[F^a(b)_i = F^a(b')_i \quad \forall i \in \overline{\Delta}(b, b')] > 1 - O(\varepsilon_a).$$

By Fact 38 we conclude that $F^a(b)$ is $O(\varepsilon_a)$-close to a direct product. ◀

Now we conclude the proof of Theorem 36. For each a, let $(g_1(a, b_1), \ldots, g_d(a, b_d))$ be a direct product that is close $O(\varepsilon_a)$-close to $F^a(b)$; the existence of these direct products was shown in Lemma 39. Then,

$$\Pr_{a,x,b}\left[f_{ab}(x) = \sum_i g_i(a, b_i)x_i\right] > 1 - O(\varepsilon). \tag{12}$$

Observe that $\phi_x(a, b) = \phi_{\mathbb{1}+x}(b, a)$. Thus, by Equation (12) and a union bound we have

$$\Pr_{a,b,x}\left[\sum_i g_i(a, b_i)x_i + f(a) = f_{ab}(x) = f_{ba}(x + \mathbb{1}) = f(b) + \sum_i g_i(b, a_i)(1 + x_i)\right] > 1 - O(\varepsilon). \tag{13}$$

For each a, b, define

$$\lambda_{ab} = \Pr_x\left[\sum_i (g_i(a, b_i) + g_i(b, a_i))x_i \neq \sum_i g_i(b, a_i) + f(a) + f(b)\right].$$

By Equation (13), $\mathbb{E}_{ab}[\lambda_{ab}] \le O(\varepsilon)$. If $\lambda_{ab} < 1/2$ then we must have

$$\sum_i g_i(b, a_i) = f(a) + f(b).$$

By Markov's inequality:

$$\Pr_{a,b}[\lambda_{ab} \ge 1/2] \le 2\mathbb{E}_{a,b}[\lambda_{ab}] \le O(\varepsilon).$$

Thus, we have:

$$\Pr_{a,b}\left[f(a) = f(b) + \sum_i g_i(b, a_i)\right] > 1 - O(\varepsilon).$$

Then, by the probabilistic method there exists β such that

$$\Pr_a\left[f(a) = f(\beta) + \sum_i g_i(\beta, a_i)\right] > 1 - O(\varepsilon).$$

That is, f is $O(\varepsilon)$-close to the direct sum $a \mapsto f(\beta) + \sum_i g_i(\beta, a_i)$. ◀

4.1 Derandomization

In this section we show how to use Theorem 36 to obtain a more randomness-efficient direct sum tester than the Square in Cube test or Diamond test. In particular, we define the ***Diamond in Cube test*** as follows: run the Diamond test on the restriction of f to a random subcube (i.e., $f(\phi_x(a, b))$).

We now show that the Direct-Sum Diamond test is more randomness-efficient than the Square in Cube test and the Diamond in Cube test for some parameter regimes. In the following discussion we assume that n is a power of two for convenience.

In the statement of tests like the Square in Cube test we say "sample $a, b \sim [n]^d, x, y \sim \mathbb{F}_2$". This would seem to indicate that the test requires $(2+2\log_2 n)d$ bits of randomness. However, if we look inside the Square in Cube test, we see that this is not the case. The Square in Cube test actually requires sampling from the following distribution:

$$(\phi_0(a, b), \phi_x(a, b), \phi_y(a, b), \phi_{x+y}(a, b)),$$

where $a, b \sim [n]^d$ and $x \sim \mathbb{F}_2^d$. However, sampling from this distribution does not actually require sampling a, b, x, y fully. Instead, we can sample from the distribution as follows. First, sample $a \sim [n]^d, x, y \sim \mathbb{F}_2^d$. Then we partially sample b. Note that we don't need to sample b_i if $x_i = y_i = 0$, because in this case the value of b_i will not be needed in any of the query points. Thus, the average number of random bits required to create a sample for the Square in Cube test is only $(2 + 1.75\log_2 n)d$.

On the surface the Diamond test might appear to be more efficient. Rather than sampling a, b, x, y it only needs to sample a, b, x. However, the Diamond test actually requires more random bits than the Square in Cube test if $n \ge 20$. The number of random bits required to sample from the Diamond in Cube's distribution is $(2\log_2 n + (1 - 1/n))d$.

Now, we claim that the Diamond in Cube test is more randomness-efficient than both the Diamond test and the Square in Cube test, at least for large n. A simple computation shows that we can obtain samples from the Diamond in Cube test's sample distribution using only $(2.5 + 1.5\log_2 n)d$ random bits, vindicating this claim.

4.2 Direct Sum Testing in the Online Adversary Model

In this section we show that an immediate application of the Affine on Subcube test from Theorem 36 yields direct sum testers that are online erasure-resilient. The t-online erasure model was first considered by Kalemaj, Raskhodnikova, and Varma in [20] and was further studied in [22, 6]. The properties that these works consider include linearity, low degree, and various properties of sequences. Let us now describe the model.

We are once again given query access to a function $f : [n]^d \to \mathbb{F}_2$, and the goal is the same as in direct sum testing — to distinguish whether f is a direct sum, or far from a direct sum. The twist is that in the t-online erasure model, there is an adversary who is allowed to erase any t entries of f after each query that the property testing algorithm makes. If a point is queried after it is erased, then the oracle will return $\perp$ instead of the actual value $f(x)$. It is not hard to see that neither the Diamond test that we analyze nor the Square in the Cube test of [13] is erasure resilient. The reason is that in both tests, after the first three queries are made, the adversary knows what the fourth query will be and can erase f at that point.

However, by using the Affine on Subcube test with an erasure resilient affinity tester, we obtain an erasure resilient direct sum tester.

▶ **Theorem 41.** *Fix a distance parameter $\delta > 0$ and erasure parameter t, and take d sufficiently large relative to n, δ, t. Then there is a direct sum tester for $f : [n]^d \to \mathbb{F}_2$ that makes $O(\max(1/\varepsilon, \log t))$-queries in the t-online erasure model that satisfies:*

- ***Completeness:*** *If f is a direct sum, the tester accepts with probability 1.*
- ***Soundness:*** *If f is δ-far from being a direct sum, then the tester rejects with probability $2/3$.*

Proof. The result follows by combining Theorem 36 with Theorem 1.1 of [6]. Specifically we use the Affine on Subcube test where the inner affinity tester is the erasure resilient one of [6]. Technically, Theorem 1.1 of [6] is stated for linearity testing, but it is straightforward to modify their result to obtain an affinity tester. ◀

5 Reconstruction for Direct Sums

In this section we address another question raised in [13] on whether a test, which they call the ***Shapka Test*** can be used to reconstruct the underlying direct sum. We show that the this is indeed the case in some parameter regimes (i.e with the input function being sufficiently close to a direct sum), and then we give an alternate method for obtaining corrected versions of the direct sum, that is both more query efficient and more error tolerant than the Shapka test. We also give upper bounds that match up to a constant for the fraction of errors that can be tolerated when reconstructing the direct sum. This shows that the new reconstruction method we propose is essentially optimal: it works for essentially the entire range of ε where recovery is information theoretically possible, and there is no asymptotically more query-efficient correction method.

We start with our lower bounds.

▶ **Proposition 42.** *Fix an even n, arbitrary integer d, and some $\epsilon > 0$ such that $n\varepsilon > 1/2$. There exist distinct direct sums $L, Z : [n]^d \to \mathbb{F}_2$, and a function $f : [n]^d \to \mathbb{F}_2$ such that* $\mathsf{dist}(f, L) = \mathsf{dist}(f, Z) \leq \varepsilon$.

Proof. Let Z be the zero function, and let L be the indicator of $x_d = 1$. Let f be any function which is zero if $x_d \neq 1$, and for inputs with $x_d = 1$, f is zero half the time. Then,

$\mathsf{dist}(f, L) = \mathsf{dist}(f, Z) = 1/(2n) \leq \varepsilon$. In other words, there is no unique closest direct sum to f, so decoding $f(\mathbb{1})$ is undefined. ◀

▶ **Proposition 43.** *Fix an n divisible by 4, an arbitrary integer d, and $\epsilon > 0$ such that $(0.7)^{d/n} < \varepsilon, n\varepsilon < 1/2$. Then there is a set $\mathcal{F}$ of functions $f : [n]^d \to \mathbb{F}_2$, such that using fewer than $n/4$ queries it is impossible to determine the value of $L(\mathbb{1})$ with probability greater than $1/2$, where L is the closest direct sum to f.*

Proof. Call a point $x \in [n]^d$ ***heavy*** if more than $2d/n$ coordinates $i \in [d]$ have $x_i = 1$. By a Chernoff bound, the fraction of points in $[n]^d$ which are heavy is less than $(0.7)^{d/n}$. Let L be a uniformly random direct sum. Define $\mathcal{F}$ to be the class of all functions which agree with L on non-heavy inputs. Then, querying $f \in \mathcal{F}$ on a heavy point is useless, because it outputs an arbitrary value unrelated to L. By our assumption that $(0.7)^{d/n} < \varepsilon$ we have that any $f \in \mathcal{F}$ is ε-close to L. Now we argue that $n/4$ non-heavy queries do not suffice to determine $L(\mathbb{1})$ with non-trivial probability. Indeed, there will be a coordinate $i \in [d]$ such that among the $n/4$ non-heavy queries, no queried point x has $x_i = 1$. Thus, we cannot hope to distinguish between the equally likely cases $L_i(1) = 0$ and $L_i(1) = 1$, and thus cannot predict $L(\mathbb{1})$ with non-trivial probability. ◀

Now we give algorithms for reconstructing values from the closest direct sum to a function f. Let us start with the ***Shapka Test*** of [13] which we now describe. This test is slightly different for odd and even d; for simplicity we only consider the case of odd d. Let $e_i \in \{0,1\}^d$ denote a vector with zeros except at the i-th coordinate. We describe how the Shapka Test can be used to reconstruct a direct sum.

▶ **Test 44** (Shapka [13])**.** Sample $a, b \in [n]^d$. Accept iff

$$f(b) = \sum_{i=1}^{d} f(\phi_{e_i}(a, b)). \tag{14}$$

Thus, to reconstruct a direct sum from a corrupted version, f, we can define,

$$f_{\mathsf{shapka}}(b) = \mathsf{maj}_{a \in [n]^d} \sum_{i=1}^{d} f(\phi_{e_i}(a, b)).$$

Dinur and Golubev show that if f passes the Shapka test with probability ε, then f is $O(n\varepsilon)$-close to a direct sum. Now, instead of checking Equation (14), we use the expression on the right hand side of Equation (14) as our guess for the value of $f(b)$, and take the most common value at each b to define the reconstructed function f_{shapka}. We show that if f is close enough to a direct sum, then f_{shapka} is in fact this direct sum:

▶ **Proposition 45.** *Fix $\varepsilon > 0, n \in \mathbb{N}$, and d odd with $n\varepsilon d < 1/4$. Suppose $f : [n]^d \to \mathbb{F}_2$ is ε-close to the direct sum $L = \sum_i L_i$. Then, for any b,*

$$\Pr_{a \in [n]^d}[f(b) = L(b)] \geq \frac{3}{4}.$$

It follows that $f_{\mathsf{shapka}} = L$.

Proof. We say that a query $f(x)$ ***succeeds*** if $f(x) = L(x)$. Let

$$\varepsilon_i = \Pr_x[f(x) \neq L(x) \mid x_i = b_i].$$

By a union bound, the Shapka test succeeds with probability at least $1 - \sum_i \varepsilon_i$. For each i,

$$\varepsilon = \Pr[f(x) \neq L(x)] \geq \Pr[f(x) \neq L(x) \wedge x_i = b_i] = \varepsilon_i / n.$$

Thus, $\sum_i \varepsilon_i \leq n\varepsilon d \leq 1/4$, so the Shapka test produces the correct value with probability at least $3/4$. ◀

Now we give a more query efficient method for reconstructing a direct sum. Our method has asymptotically optimal query complexity (by Proposition 43), and still works at essentially all possible values of ε for which the function is guaranteed to be information theoretically recoverable (by Proposition 42). We reconstruct the direct sum via the following voting scheme, where we call the reconstructed function f_{recon}. To make things more convenient, we assume that d is odd. Set

$$f_{\mathsf{recon}}(x) = \mathsf{maj}_{x^{(1)},\dots,x^{(n)}} \sum_{i=1}^{n} x^{(i)},$$

where the set of queried points $x^{(1)}, \dots, x^{(n)}$ is weighted in the majority according to the following distribution.

- Randomly partition $[d]$ into n parts, $S_1 \sqcup \cdots \sqcup S_n = [d]$, by choosing putting each $i \in [d]$ independently and uniformly at random in a part S_i.
- Sample R uniformly from $([n] \setminus \{1\})^d$.
- For $i \in [n]$, form query point $x^{(i)} \in [n]^d$ by setting $x^{(i)}_j = 1$ for each $j \in S_i$, and setting $x^{(i)}_j = R_j$ otherwise.

For even, n, we require $n+1$ queries, and use the following voting scheme.

$$f_{\mathsf{recon}}(x) = \mathsf{maj}_{x^{(1)},\dots,x^{(n)}} \sum_{i=1}^{n} x^{(i)},$$

where the set of queried points $x^{(1)}, \dots, x^{(n+1)}$ is weighted in the majority according to the following distribution.

- Randomly partition $[d]$ into $n+1$ parts, $S_1 \sqcup \cdots \sqcup S_{n+1} = [d]$, by choosing putting each $i \in [d]$ independently and uniformly at random in a part S_i.
- Sample $R \in [n]^d$ as follows: for each i independently, set $R_i = 1$ with probability $1/n^2$ and otherwise sample R_i uniformly from $[n] \setminus \{1\}$.
- For $i \in [n+1]$, form query point $x^{(i)} \in [n]^d$ by setting $x^{(i)}_j = 1$ for each $j \in S_i$, and setting $x^{(i)}_j = R_j$ otherwise.

We prove that the scheme above correctly reconstructs the underlying direct sum for the case where n is odd, as the case where n is even is similar.

▶ **Proposition 46.** *Fix n, d, ε. Let $f : [n]^d \to \mathbb{F}_2$ be ε close to a direct sum L, with $(n+1)\varepsilon < 1/4$. Then for any point x*

$$\Pr_{x^{(1)},\dots,x^{(n)}} \left[\sum_{i=1}^{n} f(x^{(i)}) = L(x) \right] \geq \frac{3}{4}.$$

It follows that $f_{\mathsf{recon}} = L$.

Proof. Without loss of generality, let us assume that $x = \mathbb{1}$ — the all ones vector. We again say that a query $f(x)$ ***succeeds*** if $f(x) = L(x)$.

We claim that $x^{(1)}, \ldots, x^{(n)}$ are uniformly random points in $[n]^d$. Fix $i \in [n]$. For each $j \in [d]$, $x_j^{(i)}$ has a $1/n$ chance of being a 1, and a $1/n$ chance of being j for any $j \in [n] \setminus \{1\}$. Furthermore, the value of $x_j^{(i)}$ is independent of the value of $x_{j'}^{(i)}$ for $j' \neq j$. Thus, the probability that all n queries all succeed is at least $1 - n\varepsilon \geq 3/4$ by a union bound. If the queries all succeed, then because d is odd, we have:

$$\sum_i f(x^{(i)}) = \sum_i L(x^{(i)}) = (d-1)L(R) + L(\mathbb{1}) = L(\mathbb{1}).$$

◀

References

1. Rudolf Ahlswede and Peter Gács. Spreading of sets in product spaces and hypercontraction of the markov operator. *The annals of probability*, pages 925–939, 1976.
2. Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron. Testing low-degree polynomials over gf (2). In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 188–199. Springer, 2003.
3. Prashanth Amireddy, Srikanth Srinivasan, and Madhu Sudan. Low-degree testing over grids. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2023)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2023.
4. Mitali Bafna, Dor Minzer, and Nikhil Vyas. Quasi-linear size pcps with small soundness from HDX. *CoRR*, abs/2407.12762, 2024. URL: `https://doi.org/10.48550/arXiv.2407.12762`, `arXiv:2407.12762`, `doi:10.48550/ARXIV.2407.12762`.
5. Mitali Bafna, Srikanth Srinivasan, and Madhu Sudan. Local decoding and testing of polynomials over grids. *Random Struct. Algorithms*, 57(3):658–694, 2020. URL: `https://doi.org/10.1002/rsa.20933`, `doi:10.1002/RSA.20933`.
6. Omri Ben-Eliezer, Esty Kelman, Uri Meir, and Sofya Raskhodnikova. Property testing with online adversaries. In Venkatesan Guruswami, editor, *15th Innovations in Theoretical Computer Science Conference, ITCS 2024, January 30 to February 2, 2024, Berkeley, CA, USA*, volume 287 of *LIPIcs*, pages 11:1–11:25. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024. URL: `https://doi.org/10.4230/LIPIcs.ITCS.2024.11`, `doi:10.4230/LIPICS.ITCS.2024.11`.
7. Arnab Bhattacharyya, Swastik Kopparty, Grant Schoenebeck, Madhu Sudan, and David Zuckerman. Optimal testing of reed-muller codes. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 488–497. IEEE, 2010.
8. Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. In *Proceedings of the twenty-second annual ACM symposium on Theory of computing*, pages 73–83, 1990.
9. Siu On Chan. Approximation resistance from pairwise-independent subgroups. *J. ACM*, 63(3):27:1–27:32, 2016. `doi:10.1145/2873054`.
10. Roee David, Irit Dinur, Elazar Goldenberg, Guy Kindler, and Igor Shinkar. Direct sum testing. In *Proceedings of the 2015 Conference on innovations in theoretical computer science*, pages 327–336, 2015.
11. Yotam Dikstein and Irit Dinur. Agreement testing theorems on layered set systems. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1495–1524. IEEE, 2019.
12. Irit Dinur. The PCP theorem by gap amplification. *J. ACM*, 54(3):12, 2007. `doi:10.1145/1236457.1236459`.
13. Irit Dinur and Konstantin Golubev. Direct Sum Testing: The General Case. In Dimitris Achlioptas and László A. Végh, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, volume

145 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 40:1–40:11, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.APPROX-RANDOM.2019.40, doi:10.4230/LIPIcs.APPROX-RANDOM.2019.40.

14 Irit Dinur and Or Meir. Derandomized parallel repetition via structured pcps. *Electron. Colloquium Comput. Complex.*, TR10-107, 2010. URL: https://eccc.weizmann.ac.il/report/2010/107, arXiv:TR10-107.

15 Irit Dinur and Omer Reingold. Assignment testers: Towards a combinatorial proof of the pcp-theorem. In *45th Symposium on Foundations of Computer Science (FOCS 2004), 17-19 October 2004, Rome, Italy, Proceedings*, pages 155–164. IEEE Computer Society, 2004. doi:10.1109/FOCS.2004.16.

16 Irit Dinur and David Steurer. Direct product testing. In *2014 IEEE 29th Conference on Computational Complexity (CCC)*, pages 188–196. IEEE, 2014.

17 Oded Goldreich and Shmuel Safra. A combinatorial consistency lemma with application to proving the PCP theorem. *SIAM J. Comput.*, 29(4):1132–1154, 2000. doi:10.1137/S0097539797315744.

18 Russell Impagliazzo, Valentine Kabanets, and Avi Wigderson. New direct-product testers and 2-query pcps. *SIAM J. Comput.*, 41(6):1722–1768, 2012. doi:10.1137/09077299X.

19 Jeff Kahn, Gil Kalai, and Nathan Linial. *The influence of variables on Boolean functions.* Institute for Mathematical Studies in the Social Sciences, 1989.

20 Iden Kalemaj, Sofya Raskhodnikova, and Nithin Varma. Sublinear-time computation in the presence of online erasures. In Mark Braverman, editor, *13th Innovations in Theoretical Computer Science Conference, ITCS 2022, January 31 - February 3, 2022, Berkeley, CA, USA*, volume 215 of *LIPIcs*, pages 90:1–90:25. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022. URL: https://doi.org/10.4230/LIPIcs.ITCS.2022.90, doi:10.4230/LIPICS.ITCS.2022.90.

21 Tali Kaufman and Madhu Sudan. Algebraic property testing: the role of invariance. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 403–412, 2008.

22 Dor Minzer and Kai Zhe Zheng. Adversarial low degree testing. In David P. Woodruff, editor, *Proceedings of the 2024 ACM-SIAM Symposium on Discrete Algorithms, SODA 2024, Alexandria, VA, USA, January 7-10, 2024*, pages 4395–4409. SIAM, 2024. doi:10.1137/1.9781611977912.154.

23 Dor Minzer and Kai Zhe Zheng. Near optimal alphabet-soundness tradeoff pcps. In Bojan Mohar, Igor Shinkar, and Ryan O'Donnell, editors, *Proceedings of the 56th Annual ACM Symposium on Theory of Computing, STOC 2024, Vancouver, BC, Canada, June 24-28, 2024*, pages 15–23. ACM, 2024. doi:10.1145/3618260.3649606.

24 Ryan O'Donnell. Analysis of boolean functions, 2021. URL: https://arxiv.org/abs/2105.10386, arXiv:2105.10386.

25 Ran Raz and Shmuel Safra. A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In Frank Thomson Leighton and Peter W. Shor, editors, *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing, El Paso, Texas, USA, May 4-6, 1997*, pages 475–484. ACM, 1997. doi:10.1145/258533.258641.