THEORY OF COMPUTING, Volume 19 (3), 2023, pp. 1–56
www.theoryofcomputing.org

Exponentially Small Soundness for the Direct Product Z-test

Irit Dinur* Inbal Livni Navon†

Received July 4, 2019; Revised October 12, 2022; Published October 2, 2023

Abstract. Given a function $f : [N]^k \to [M]^k$, the Z-test is a three-query test for checking if the function f is a direct product, i. e., if there are functions $g_1, \ldots, g_k : [N] \to [M]$ such that $f(x_1, \ldots, x_k) = (g_1(x_1), \ldots, g_k(x_k))$ for every input $x \in [N]^k$.

This test was introduced by Impagliazzo et. al. (SICOMP 2012), who showed that if the test passes with probability $\epsilon > \exp(-\sqrt{k})$ then f is $\Omega(\epsilon)$-correlated to a direct product function in some precise sense. It remained an open question whether the soundness of this test can be pushed all the way down to $\exp(-k)$ (which would be optimal). This is our main result: we show that whenever f passes the Z-test with probability $\epsilon > \exp(-k)$, there must be a global reason for this, namely, f is $\Omega(\epsilon)$-correlated to a direct product function, in the same sense of closeness.

Towards proving our result we analyze the related (two-query) V-test, and prove a "restricted global structure" theorem for it. Such theorems were also proven in previous work on direct product testing in the small soundness regime. The most recent paper, by Dinur and Steurer (CCC 2014), analyzed the V-test in the exponentially small soundness regime. We strengthen their conclusion by moving

A preliminary version of this paper appeared in the Proceedings of the 32nd Computational Complexity Conference, 2017 [7].

*Research supported in part by an ISF-UGC grant number 1399/14, and by BSF grant number 2014371.

†Research supported in part by an ISF-UGC grant number 1399/14, and by BSF grant number 2014371.

ACM Classification: F.2.2

AMS Classification: 68Q25

Key words and phrases: direct product testing, property testing, agreement

© 2023 Irit Dinur and Inbal Livni Navon
Licensed under a Creative Commons Attribution License (CC-BY) DOI: 10.4086/toc.2023.v019a003

from an "in expectation" statement to a stronger "concentration of measure" type of statement, which we prove using reverse hypercontractivity. This stronger statement allows us to proceed to analyze the Z-test.

1 Introduction

A function $f : [N]^k \to [M]^k$ for $N, M, k \in \mathbb{N}$ is a *direct product function* if $f = (g_1, \ldots, g_k)$, for $g_i : [N] \to [M]$, i. e., the output of f on each coordinate depends on the input to this coordinate alone. Direct products appear in a variety of contexts in complexity theory, usually for hardness amplification. In PCP constructions it underlies the Parallel Repetition Theorem [18] and implicitly appears in other forms of gap amplification, see e. g., [4]. The specific task of testing direct products as an abstraction of a certain element of PCP constructions was introduced by [9].

The combinatorial question that underlies these constructions is the direct product testing question: given a function $f : [N]^k \to [M]^k$, is it a direct product function? The setting of interest here is where we query f on *as few inputs as possible*, and decide if is it a direct product function. Direct product testing is a type of property testing question, yet it is not in the standard property testing parameter regime. In property testing, we are generally interested in showing that functions that pass the test with high probability, for example 99%, are close to having the property.

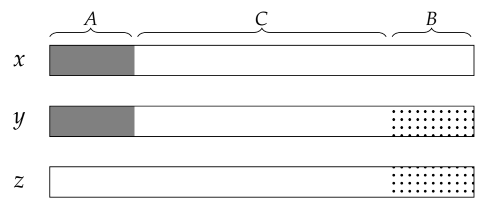
In our case, we are interested in understanding the structure of functions that pass the test with small—but non-trivial—probability, e. g., 1%. The 1% regime is often more challenging than the 99% regime. It plays an important role in PCPs where one needs to prove a large gap. In such arguments, one needs to be able to deduce non-trivial structure even from a proof that passes a verification test with small probability, e. g., 1%.

There are very few families of tests for which 1% theorems are known. These include algebraic low-degree tests and direct product tests. For low-degree tests there has been a considerable amount of work in various regimes and in particular towards understanding the extent of the 1% theorems, see, e. g., [19, 1, 3] and [2]. It is intriguing to understand more broadly for which tests such theorems can hold. Indeed, as far as we know, there are no other tests that exhibit such strong "structure vs. randomness" behavior, and direct product tests are natural candidates in which to study this question.

We remark that finding new settings where 1% theorems hold (including in particular derandomized direct products) can potentially be useful for constructing locally testable codes and stronger PCPs, see, e. g., the recent papers [14, 6]. Towards this goal, gaining a more comprehensive understanding of direct product tests, as well as developing tools for proving them, are natural objectives. Our results improve the minimal soundness of the 2-query PCP from [13] from $\exp(\sqrt{k})$ to $\exp(k)$.

Test 1: Z-test with parameter t (3-query test)

1. Choose A, B, C to be a random partition of $[k]$, such that $|A| = |B| = t$.
2. Choose uniformly at random $x, y, z \in [N]^k$ such that $x_A = y_A$ and $y_B = z_B$.
3. Accept if $f(x)_A = f(y)_A$ and $f(z)_B = f(y)_B$.



Denote by $agr_t^Z(f)$ *the success probability of* f *on this test.*

1.1 Our main result

The main question we study is: if $f : [N]^k \to [M]^k$ passes a certain natural test (Test 1) with non-negligible probability, what does f look like?

Theorem 1.1 (Main Theorem – Global Structure)**.** *For every* $N, M > 1$, *there exists* $c > 0$ *such that for every* $\lambda > 0$ *and large enough* k, *if* $f : [N]^k \to [M]^k$ *is a function that passes Test 1 with probability* $agr_{k/10}^Z(f) = \epsilon \geq e^{-c\lambda^2 k}$, *then there exist functions* $(g_1, \ldots, g_k)$, $g_i : [N] \to [M]$ *such that*

$$\Pr_{x \in [N]^k}\left[f(x) \overset{\lambda k}{\approx} (g_1(x_1) \ldots, g_k(x_k)) \right] \geq \frac{\epsilon}{10},$$

where $\overset{\lambda k}{\approx}$ *means that the strings are equal on all but at most* λk *coordinates.*

The theorem is qualitatively tight with respect to several parameters: (i) soundness (i. e., the parameter ϵ), (ii) approximate equality vs. exact equality (i. e., the parameter λ), (iii) number of queries in the test. We discuss these next.

(i) Soundness The soundness of the theorem is the smallest success probability for which the theorem holds. In our case it is 2^{-ck} for some constant $c > 0$. This is tight up to the constant c, as can be seen from the example below.

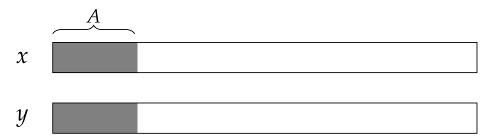
Example 1.2 (Random function)**.** Let $f : [N]^k \to \{0, 1\}^k$ be a random function, i. e., for each $x \in [N]^k$ choose $f(x) \in \{0, 1\}^k$ uniformly and independently. Two random strings in $\{0, 1\}^t$ are equal with probability 2^{-t}, therefore $\mathrm{agr}_t^Z(f) \geq 2^{-2t}$, since the test performs two such checks. On the other hand, since f is random, it is not close to any direct product function (see Section 6 for more information).

We remark that every function $f : [N]^k \to \{0, 1\}^k$ is at least 2^{-k} close to a direct product function [1], so this amount of correlation is meaningless. We conclude that in order to have direct product theorem that is not trivial, the minimal soundness has to be larger than 2^{-k}.

[1]Consider the direct product function constructed incrementally by taking the most common value out of $\{0, 1\}$ on each step.

Test 2: V-test with parameter t (2-query test)

1. Choose $A \subset [k]$ of size t, uniformly at random.
2. Choose uniformly at random $x, y \in [N]^k$ such that $x_A = y_A$.
3. Accept if $f(x)_A = f(y)_A$.



Denote by $agr_t^{\vee}(f)$ the success probability of f on this test.

(ii) Approximate equality vs. exact equality In the theorem, we prove that for for an $\Omega(\epsilon)$ fraction of the inputs x we have $f(x) \overset{\lambda k}{\approx} (g_1(x), \dots, g_k(x))$. A priori, one could hope for a stronger conclusion in which $f(x) = (g_1(x), \dots, g_k(x))$ for an $\Omega(\epsilon)$ fraction of the inputs x. However, Example 1.3 shows that for $t = k/10$, approximate equality is necessary.

Example 1.3 (Noisy direct product function)**.** This example is from [5]. Let f be a direct product function, except that on each input x we "corrupt" $f(x)$ on λk random coordinates by changing $f(x)$ on these coordinates into random values. For $\lambda > 1/10$, the probability that Test 1 on f misses all the corrupted coordinates is $2^{-\Omega(\lambda k)}$, in which case the test succeeds. Since we have changed $f(x)$ on λk coordinates into random values, no direct product function can approximate f on more than a $(1 - \lambda)$ fraction of the coordinates.

From this example, we conclude that it is not possible to approximate f that passes Test 1 (with parameter $t = k/10$) with probability $\mathrm{e}^{-c\lambda k}$ on more than a $(1-\lambda)$ fraction of the coordinates. In Section 6 we prove similar bounds for different intersection sizes, and also discuss different test variants.

(iii) Number of queries in the test The absolute minimum number of queries for any direct product test is two. Indeed, there is a very natural 2-query test, Test 2. Dinur and Goldenberg showed that it is not possible to have a direct product theorem with soundness lower than $1/\mathrm{poly}(k)$ using the 2-query test [5].

Example 1.4 (Localized direct product functions)**.** In this example we assume that $N = \omega(k^2)$. For every $b \in [N]$ we choose a random function $g_b : [N] \to [M]$ independently. For every input $x \in [N]^k$, we choose a random $i_x \in k$, set $b = x_i$ and set $f(x) = (g_b(x_1), \dots, g_b(x_k))$.

The function f satisfies $\mathrm{agr}_t^{\vee}(f) \geq t/k^2$; indeed, for x, y and A chosen in the test, if $i_x = i_y$ and $i_x \in A$, then the test will pass. The probability that $i_x = i_y$ is $1/k$, and the probability that $i_x \in A$ is t/k.

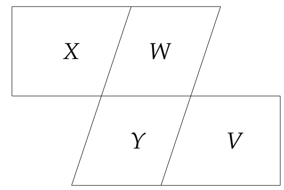
For $N = \omega(k^2)$, the function f is far from direct product, since it is made up from N different direct product functions, each piece consisting of roughly a $1/N$ fraction of the domain $[N]^k$.

For every t, the function described in the example satisfies $\mathrm{agr}_t^{\vee}(f) \geq 1/k^2$, yet there is no direct product function that approximates f on $\Omega(1/k^2)$ fraction of the domain. In [5] the

Test 3: Z-test for functions over sets, with parameter t (3-queries)

1. Choose random $V, W, X, Y \subset [N]$, such that $|W| = |V| = t$, $|X| = |Y| = k - t$ and $X \cap W = Y \cap W = Y \cap V = \emptyset$.

2. Accept if $f(X \cup W)_W = f(Y \cup W)_W$ and $f(Y \cup W)_Y = f(Y \cup V)_Y$.



Denote by $agr_t^{Z_{set}}(f)$ the success probability of f on this test.

conclusion from Example 1.4 was that $1/\mathrm{poly}(k)$ is the limit for small soundness for direct product tests. However, [13] showed that by adding just one more query, this limitation goes away. They introduced a 3-query test, similar to Test 1, and proved a direct product theorem for all $\epsilon > 2^{-k^{\beta}}$ for some constant $\beta \leq 1/2$.

Direct product test for functions over sets Some of the previous results on direct products, such as [13], were proven in a slightly different setting where the function tested is $f : \binom{[N]}{k} \to [M]^k$. The input to f is an unordered set $S \subset [N]$ of k elements, and we view $f(S)$ a matching that matches each element $a \in S$ to an element $f(S)_a \in [M]$. The first bit of $f(S)$ corresponds to $f(S)_a$ for the smallest elements $a \in S$, and so on. In this setting, a direct product function is $g : [N] \to [M]$, and we say that $f(S) \overset{t}{\approx} g(S)$ if for all but t of the elements $a \in S$, $f(S)_a = g(a)$.

In this article, we prove a direct product testing theorem also for this setting. Test 3 is the analog of Test 1 for functions over sets. In Test 3 (see figure), we pick four sets $W, X, Y, V \subset [N]$, such that $X \cap W = Y \cap W = Y \cap V = \emptyset$ (other intersections can be non-empty). The sets are picked such that $|X \cup W| = |Y \cup W| = |Y \cup V| = k$, so that they can be inputs for f.

Theorem 1.5 (Global Structure for Sets)**.** *There exists a constant $c > 0$, such that for every $\lambda > 0$, large enough $k \in \mathbb{N}$ and $N > \mathrm{e}^{c\lambda k}$, $M \in \mathbb{N}$, if the function $f : \binom{[N]}{k} \to [M]^k$ passes Test 3 with probability $agr_{k/10}^{Z_{set}}(f) = \epsilon > \mathrm{e}^{-c\lambda k}$, then there exists a function $g : [N] \to [M]$ such that*

$$\Pr_S \left[f(S) \overset{\lambda k}{\approx} g(S) \right] \geq \epsilon - 4\epsilon^2 .$$

Notice that the bound $\epsilon - 4\epsilon^2$ is better than the bound in Theorem 1.1, and it is tight, as demonstrated by the function f which is a hybrid of $1/\epsilon$ different direct product functions on equal parts of the inputs. Such function f passes Test 3 with probability ϵ, and every direct product function is close to f only on an ϵ fraction of the inputs.

We remark that the two theorems are not the same. In Theorem 1.1, there are k different functions $g_1, \ldots, g_k : [N] \to [M]$ whereas in Theorem 1.5 there is a single one. Furthermore, Theorem 1.1 holds for any $N, M \in \mathbb{N}$ and large enough k, and Theorem 1.5 (and other such direct product theorems) only hold for $N \gg k$. The proofs of the theorems are also different, as discussed later in the Introduction.

The proof of the global structure for sets uses the fact that the first query, $X \cup W$, and the last query, $Y \cup V$, are nearly independent. These queries are not completely independent, because Y, W are picked such that $Y \cap W = \emptyset$. The difference between the distribution of $X \cup W$ and $Y \cup V$ and the distribution of two independent subsets of size k is bounded by k^2/N. This means that the theorem holds when $k^2/N \ll \epsilon$. For an exponentially small ϵ, this implies that $N = \exp(k)$. We have not analyzed the case where N is smaller with respect to k in this setting (in the main theorem, when $f : [N]^k \to [M]^k$, there is no requirement that N should be large with respect to k).

1.2 Restricted global structure

Our proof has two main parts, similar to the structure of the proof of [5, 13]. In the first part, we analyze only Test 2 and prove a restricted global structure theorem for it, Theorem 1.6 below, (this was called local structure in [13, 8]). The term "restricted global structure" refers to the case when we restrict the domain to small (but not trivial) pieces, and show that f is close to a product function on each piece separately. This is the structure of the function in Example 1.4.

More explicitly, for every $A \in [k]$ of size $k/10$, $r \in [N]^A$ and $\gamma \in [M]^A$, a *restriction* is a triple $\tau = (A, r, \gamma)$. The choice of $t = k/10$ in Theorem 1.1 is somewhat arbitrary, the theorem can be proven with $t = ck$ for any constant $c < 1/2$. The restriction corresponds to the set of inputs

$$\mathcal{V}_\tau = \{w \in [N]^{[k]\setminus A} | f(r, w)_A = \gamma\}.$$

Our restricted global structure theorem is that for many restrictions τ, there exists a direct product function that is close to f on $\mathcal{V}_\tau$.

Theorem 1.6 (Restricted Global Structure – informal)**.** *There exists a constant $c > 0$, such that for every $\alpha > 0$ and large enough $k \in \mathbb{N}$ the following holds. Let $f : [N]^k \to [M]^k$ be a function that passes Test 2 with probability $agr^{\vee}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\alpha k}$. Let $\mathcal{D}$ be the test distribution over τ, namely choosing $A \subset [k]$, $x \in [N]^k$ uniformly and setting $\tau = (A, x_A, f(x)_A)$. Then with probability $\Omega(\epsilon)$, there exists a direct product function $g = (g_1, \dots, g_{9k/10})$, $g_i : [N] \to [M]$ such that,*

$$\Pr_{w \in [N]^{[k]\setminus A}} \left[f(r, w)_{[k]\setminus A} \overset{\alpha k}{\approx} g(w_{[k]\setminus A}) \,\middle|\, w \in \mathcal{V}_\tau \right] \geq 1 - \epsilon^2 \tag{1.1}$$

A similar theorem was proven in [13] but only for soundness (i. e., ϵ) at least $\exp(-k^\beta)$ for a constant $\beta \leq 1/2$. This was strengthened to soundness $\exp(-\Omega(k))$ in [8]. Our Theorem 1.6 improves on the conclusion of [8]. In [8] the probability in (1.1) was shown to be at least $1 - O(\alpha)$ (recall that α is a constant), whereas we show it is exponentially close to 1 (when ϵ is that small). This difference may seem minor but in fact it is what prevented [8] from deriving global structure via a three-query test (i. e., moving from the V-test to the Z-test). When we try to move from restricted global structure to global structure, the consistency inside each restriction needs to be very high for the probabilistic arguments to work, as we explain below.

The restricted global structure gives us a direct product function that approximates f only on a restricted subset of the inputs. In the proof of the global structure, we use the third query to show that there exists a global function. A key step in the proof of the global structure is to show that for many restrictions τ, the function g_τ is close to f on a much larger subsets of inputs. This is done, intuitively, by claiming that if $f(x)_A = f(y)_A$, then with high probability $f(y) \approx g_\tau(y)$ for $\tau = (A, x_A, f(x)_A)$. Since B is a random set and $f(z)_B = f(y)_B$, then $f(z), g_\tau(z)$ are also close. This claim only holds if the success probability on (1.1) is more than $1 - \epsilon$, else it is possible that all the success probability of the test comes from f such that $f(x)_A = f(y)_A$, but $f(y), g_\tau(y)$ are far from each other.

1.3 Technical contribution

In terms of technical contributions our proof consists of two new components.

Domain extension Our first contribution is a new *domain extension* step that facilitates the proof of the restricted global structure. The restricted global structure shows that with probability $\Omega(\epsilon)$, the function f is close to a direct product function on the restricted domain $\mathcal{V}_\tau$. A natural way to show that a function is close to a direct product function is to define a direct product function by majority value. However, this method fails when the agreement guaranteed for f is small, as in our case.

This is usually resolved by moving to a restricted domain in which the agreement is much higher, and by defining majority there. The first part of our proof is to show that with probability $\Omega(\epsilon)$ over the restrictions $\tau = (A, r, \gamma)$, the set $\mathcal{V}_\tau$ satisfies the two properties:

1. Its density is at least $\epsilon/2$.
2. f has very high agreement in $\mathcal{V}_\tau$. Informally it means that taking a random pair $w, v \in \mathcal{V}_\tau$ which agree on a random subset of coordinates J, then $f(r, w)_J \approx f(r, v)_J$ with probability greater than $1 - \epsilon$.

We call such restrictions excellent, following [13].

We show that for every excellent restriction $\mathcal{V}_\tau$, f is close to a direct product function on $\mathcal{V}_\tau$. Let $f_\tau : \mathcal{V}_\tau \to [M]^{k\setminus A}$ be the restriction of f to $\mathcal{V}_\tau$, i. e., $\forall w \in \mathcal{V}_\tau, f_\tau(w) = f(r, w)_{[k]\setminus A}$. The function f_τ has high agreement, which is good for defining majority, but unfortunately $\mathcal{V}_\tau$ is very sparse in $[N]^{k\setminus A}$. The density of $\mathcal{V}_\tau$ can be as low as $\epsilon/2$, which is exponentially small, and this is where the techniques used in [13] break down. In order to prove that f is close to a direct product function on $\mathcal{V}_\tau$, we use a local averaging operator to *extend* the domain from $\mathcal{V}_\tau$ to $[N]^{[k]\setminus A}$.

The local averaging operator $\mathcal{P}_{\frac{3}{4}}$ is the majority of a 3/4-correlated neighborhood of $\mathcal{V}_\tau$,

$$\forall w \in [N]^{[k]\setminus A}, i \in [k] \setminus A \qquad \mathcal{P}_{\frac{3}{4}} f_\tau(w)_i = \operatorname*{Plurality}_{v \underset{3/4}{\sim} w, v\in\mathcal{V}_\tau, v_i = w_i} \{f_\tau(v)_i\},$$

where $v \underset{3/4}{\sim} w$ means that for every $j \neq i$ independently, we take $v_j = w_j$ with probability $1/4$, and a random value in $[N]$ else.

The domain of the new function is all of $[N]^{[k]\setminus A}$, $\mathcal{P}_{\frac{3}{4}} f_\tau : [N]^{[k]\setminus A} \to [M]^{[k]\setminus A}$, and it satisfies the following properties:

1. $\mathcal{P}_{\frac{3}{4}} f_\tau$ approximates f_τ on $\mathcal{V}_\tau$.

2. $\mathcal{P}_{\frac{3}{4}} f_\tau$ has high agreement, taking a random pair $w, v \in [N]^{[k]\setminus A}$ such that $w_J = v_J$, results in agreeing answers, $\mathcal{P}_{\frac{3}{4}} f_\tau(w)_J \approx \mathcal{P}_{\frac{3}{4}} f_\tau(v)_J$ with probability greater than $1 - \epsilon$.

The main technical tool we use to prove these properties is a reverse hypercontractivity argument from [16]. For every two sets $U, V \subset [N]^k$, Mossel et al. proved a lower bound on the probability of a random $w \in [N]^k$ and $v \sim_{3/4} w$ to be in U and in V, respectively. We use their result to prove that for almost all of $w \in [N]^{[k]\setminus A}$, taking $v \underset{3/4}{\sim} w$ ends inside $\mathcal{V}_\tau$ with probability at least poly(ϵ). This allows us to prove that for almost all of $w \in [N]^{[k]\setminus A}, i \in [k] \setminus A$, the plurality $\mathcal{P}_{\frac{3}{4}} f_\tau(w)_i$ relays on *many values* of $f_\tau(v)_i$, which in turn lets us to prove the two properties above.

Lastly, we define a direct product function g_τ by taking the plurality over $\mathcal{P}_{\frac{3}{4}} f_\tau$, and show that it is close to f_τ.

Direct product testing in a dense regime A second new element comes when stitching the many localized functions into one global direct product function, by using the third query.

We prove two global structure theorems, Theorem 1.1 for functions on tuples (ordered lists) $f : [N]^k \to [M]^k$ and Theorem 1.5 for functions on sets $f : \binom{[N]}{k} \to [M]^k$.

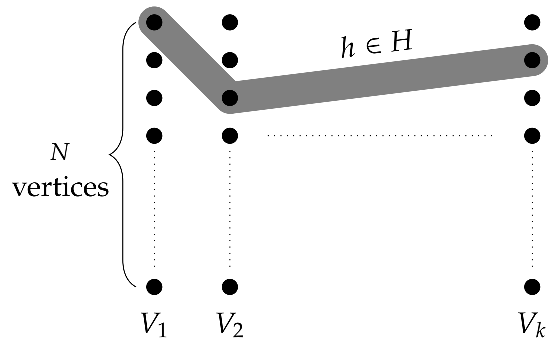
When we work with f that is defined over sets, we can directly follow the approach of [13] to complete the proof. However, when working with f defined on tuples we reach a combinatorial question that itself resembles a direct product testing question, but in a different (dense) regime. Luckily, the fact that this question is in a dense regime makes it easier to solve, and this leads to our global structure theorem for tuples.

1.4 Agreement tests and direct product tests

The question of direct product testing fits into a more general family of tests called agreement tests. We digress slightly to describe this setting formally and explain how direct product tests fit into this framework.

Agreement tests In all efficient PCPs we break a proof into small overlapping pieces, use relatively inefficient PCPs (i. e., PCPs that incur a large blowup) to encode each small piece, and then through an *agreement test* put the pieces back together. The agreement test is needed because given the set of pieces (their values and locations), there is no guarantee that the different pieces come from the same underlying global proof, i. e., that the proofs of each piece can be "put back together again". The PCP system needs to ensure this through *agreement testing*: we take two (or more) pieces that have some overlap, and check that they agree.

Figure 1: complete k-uniform k-partite graph



This situation can be formulated as an agreement testing question as follows. Let V be a ground set, $|V| = N$, and let H be a collection of subsets of V, i. e., a set of hyperedges. Let $[M]$ be a finite set of colors, where it is sufficient to think of $M = 2$.

A *local assignment* is a set $a = \{a_s\}$ of local colorings $a_s : s \longrightarrow [M]$, one per subset $s \in H$. A local assignment is called *global* if there is a global coloring $g : V \longrightarrow [M]$ such that

$$\forall s \in H, \qquad a_s \equiv g|_s .$$

An *agreement check* for a tuple of subsets $s_1, \dots, s_q$ checks whether their local functions agree on any point in the intersection, denoted $\mathrm{agree}(a_{s_1}, \dots, a_{s_q})$. Formally,

$$\mathrm{agree}(a_{s_1}, \dots, a_{s_q}) \qquad \Leftrightarrow \qquad \forall i, j \in [q],\ x \in s_i \cap s_j, \quad a_{s_i}(x) = a_{s_j}(x) .$$

A local assignment that is global passes all agreement checks. The converse is also true: a local assignment that passes *all* agreement checks must be global.

An *agreement test* is specified by giving a distribution $\mathcal{D}$ over tuples of subsets $s_1, \dots, s_q$. We define the agreement of a local assignment to be the probability of agreement,

$$\underset{\mathcal{D}}{\mathrm{agr}}(a) = \Pr_{(s_1,\dots,s_q)\sim\mathcal{D}} \left[\mathrm{agree}(a_{s_1}, \dots, a_{s_q})\right] .$$

An agreement theorem shows that if a is a local assignment with $\mathrm{agr}_{\mathcal{D}}(a) > \epsilon$ then a is somewhat close to a global assignment. Agreement theorems can be studied for any hypergraph and in this article we prove such theorems for two specific hypergraphs: the k-uniform complete hypergraph, and the k-uniform k-partite complete hypergraph.

Relation to direct product testing Theorem 1.5 is readily interpretted as an agreement test theorem. As for Theorem 1.1, we next describe a hypergraph on which it can also be interpreted "geometrically" as an agreement test theorem. Consider *complete k-uniform k-partite hypergraph* (see Figure 1). Let $G = (V = V_1, \dots, V_k, H)$ be the complete k-partite hypergraph with $|V_i| = N$ for $i \in [k]$, and

$$H = \{(v_1, \dots, v_k) \mid \ \forall i \in [k], v_i \in V_i\} .$$

There is a bijection between H and $[N]^k$. We shall interpret $f(x_1,\ldots,x_k)$ as a local coloring of the vertices $x_1,\ldots,x_k$. In this way, we have the following equivalence

$$f:[N]^k\to[M]^k \qquad\Longleftrightarrow\qquad a=\{a_x\}_{x\in H}.$$

Moreover, local assignments which are global, i. e., a such that $a_x = g|_x$ for some global coloring $g: V_1\cup\cdots\cup V_k\to[M]$, correspond exactly to functions f which are direct products, $f=(g_1,\ldots,g_k)$ where $g_i = g|_{V_i}$,

$$f=(g_1,\ldots,g_k) \qquad\Longleftrightarrow\qquad a \text{ is global.}$$

Finally, Test 2 can be described as taking 2 hyperedges that intersect on t vertices, and check if their local functions agree on the intersection. Similarly, Test 1 can be described as picking three hyperedges, $h_1,h_2,h_3\in H$ such that h_1,h_2 intersect on t vertices, and h_2,h_3 intersect on a disjoint set of t vertices, and checking agreement.

Our main theorem, Theorem 1.1, is equivalent to an agreement theorem showing that if a local assignment a passes a certain 3-query agreement test with non-negligible probability, then there exists a global assignment $g: V\to[M]$ with which it agrees non-negligibly.

The k-uniform complete hypergraph (it is non-partite, in contrast to the above), is related to Theorem 1.5. In this hypergraph the vertex set is $[N]$ and there is a hyperedge for every possible k-element subset of $[N]$. Now we have a similar equivalence between local assignments and functions over sets, i. e., functions where the input is a set $S\subset[N]$ of size k,

$$f:\binom{[N]}{k}\to[M]^k \qquad\Longleftrightarrow\qquad a=\{a_s\}_{s\in\binom{[N]}{k}}.$$

An agreement theorem for this hypergraph is equivalent to Theorem 1.5, in which f is defined not on the set of tuples $[N]^k$ but on the set of subsets $\binom{[N]}{k}$. A global assignment a on this graph is equivalent to a direct product function over sets, i. e., $f=g:[N]\to[M]$.

1.5 Organization of the paper

Section 2 contains preliminary notation and definitions. In Section 3 we prove the restricted global structure, Theorem 1.6. Section 4 is dedicated to the global structure for functions on sets. We show how to deduce a variant of Theorem 1.6 for sets rather than tuples and then prove the global structure theorem for sets, Theorem 1.5. In Section 5 we prove the global structure theorem for tuples, Theorem 1.1. Lastly, in Section 6 we discuss lower bounds for various 3-query direct product tests that were not presented in the introduction.

The conference version of this paper had a small gap in the proof that has been thankfully caught by the careful reviewers and editor, and this has been corrected in Appendix A by referencing to [12].

2 Preliminaries

For a set $A\subset[k]$ we denote by $\bar{A}$ the set $[k]\setminus A$.

Definition 2.1. For every 2 strings $x, y \in [N]^k$ we say that

1. $x \overset{t}{\approx} y$ if x, y differ in at most t coordinates.

2. $x \overset{t}{\not\approx} y$ if x, y differ in more than t coordinates.

Notice that the distance in the above definition is not relative, but it is the number of coordinates in which the two strings differ.

For a set $A \subset [k]$, we denote by $r \in [N]^A$ a matching from every $i \in A$ to an element in $[N]$. For $r \in [N]^A$ and $w \in [N]^{\bar{A}}$, the string $(r, w) \in [N]^k$ is the string created by taking for each $i \in A$ the element matched to i in r, and for each $i \notin A$ the element matched to i in w.

Let $x \in [N]^k$, we denote by $x_A \in [N]^A$ the matching which matches each $i \in A$ the element $x_i \in [N]$.

For $x, y \in [N]^A$, we say that $x \overset{t}{\approx} y$ if for all except at most t coordinates $i \in A$, x, y match the same value to i.

Definition 2.2 (Plurality)**.** The plurality of a function f on a distribution $\mathcal{D}$ is its most frequent value

$$\underset{x\sim\mathcal{D}}{\mathrm{Plurality}}(f(x)) = \arg\max_{\beta} \left\{ \Pr_{x\in\mathcal{D}}[f(x) = \beta] \right\} .$$

We use a few different Bernstein-Chernoff-type concentration bounds, stated below.

Fact 2.3 (Chernoff bound)**.** *Let $X_1, \ldots, X_k$ be independent random variables in $\{0, 1\}$, let $X = \sum_{i=1}^{k} X_i$ and denote $\mu = \mathbb{E}[X]$. Then for every $\delta \in (0, 1)$,*

$$\Pr_{X_1,\ldots,X_k}[X - \mu \geq \delta\mu] \leq \mathrm{e}^{-\frac{\delta^2\mu}{3}} \tag{2.1}$$

$$\Pr_{X_1,\ldots,X_k}[X - \mu \leq -\delta\mu] \leq \mathrm{e}^{-\frac{\delta^2\mu}{2}} . \tag{2.2}$$

Inequality (2.1) appears as [10, Eq. (6)] and [15, Thm. 4.4.2]. Inequality (2.2) appears as [10, Eq. (7)] and [15, Thm. 4.5.2].

We use two variants of this bound for sampling without replacement. The next variant follows from combining Fact 2.3, Eq. (2.2) with Hoeffding's generic reduction from sampling without replacement to a sum of independent variables, [11, Theorem 4].

Fact 2.4 (Hoeffding bound for random subset.)**.** *Let C be a set, and let $B \subset C$ be subset. Suppose we pick a random subset $S \subset C$ of size $n < |C|$, then for every $\delta \in (0, 1)$*

$$\Pr_{S}\left[|B \cap S| \leq (1-\delta)\frac{|B|n}{|C|}\right] \leq \mathrm{e}^{-\frac{|B|n\delta^2}{2|C|}} .$$

We also use the following result by Hoeffding.

Fact 2.5 (Hoeffding's inequality for random sampling without replacement, [11, Theorems 1 and 4])**.** *Let $C = \{c_1, \dots c_k\}$ be a multiset of k values, $c_i \in [0,1]$, and let $X_1, \dots, X_n$ be n random samples without replacement from C. Let $X = \sum_{i=1}^{k} X_i$ and denote $\mu = \mathbb{E}[X]$, then*

$$\Pr_{X_1,\dots,X_k}[X - \mu > t] \le \mathrm{e}^{-\frac{2t^2}{n}}.$$

The random variables in Hoeffding's bounds for random variables without replacement are not independent, but they are not independent in a very specific way, of being without replacement. In addition to these bounds we also have the following bound for random variables that are not independent. This bound appears as Theorem 1.1 in [12].

Fact 2.6 (Generalized Chernoff bound [12, 17])**.** *Let $X_1, \dots, X_k$ be random variables in $\{0,1\}$, let $X = \sum_{i=1}^{k} X_i$. If there exists $\eta \in (0,1)$ such that for every $S \subseteq [k]$, $\Pr[\wedge_{i \in S} X_i = 1] \le \eta^{|S|}$, then for every $\gamma \in (\eta, 1)$,*

$$\Pr_{X_1,\dots,X_k}[X \ge \gamma k] \le \mathrm{e}^{-2k(\gamma-\eta)^2}.$$

2.1 Reverse hypercontractivity

Definition 2.7 (ρ-correlated distribution)**.** For every string $x \in [N]^k$ and constant $\rho \in (0,1)$, the ρ-correlated distribution from x, denoted by $y \underset{\rho,J}{\sim} x$, is defined as follows. Each $i \in [k]$ is inserted into J with probability ρ, independently. The string y is chosen such that $x_J = y_J$, and the rest is uniform. In some cases, we omit the subscript J.

We quote Proposition 9.2 from [16]:

Proposition 2.8. *Let $A, B \subseteq [N]^k$ of sizes $\Pr_{w \in [N]^k}[w \in A] = \mathrm{e}^{-\frac{a^2}{2}}$ and $\Pr_{w \in [N]^k}[w \in B] = \mathrm{e}^{-\frac{b^2}{2}}$. Then*

$$\Pr_{x \in [N]^k, y \underset{\rho}{\sim} x}[x \in A, y \in B] \ge \mathrm{e}^{-\frac{(2-\rho)(a^2+b^2)}{4(1-\rho)} - \frac{\rho ab}{2(1-\rho)}}.$$

By changing notation and simplifying, we get the following corollary.

Corollary 2.9. *For $A, B \subseteq [N]^k$, $|A| \ge |B|$,*

$$\Pr_{x \in [N]^k, y \underset{\rho}{\sim} x}[x \in A, y \in B] \ge \Pr_{x \in [N]^k}[x \in A]^{1+\frac{\rho}{2(1-\rho)}} \Pr_{x \in [N]^k}[x \in B]^{1+\frac{3\rho}{2(1-\rho)}}.$$

Proof. $|A| \ge |B|$ implies $a \le b$, we know that

$$\mathrm{e}^{-\frac{\rho ab}{2(1-\rho)}} \ge \mathrm{e}^{-\frac{\rho b^2}{2(1-\rho)}} = \Pr_{x \in [N]^k}[x \in B]^{\frac{\rho}{1-\rho}}.$$

Similarly

$$e^{-\frac{(2-\rho)(a^2+b^2)}{4(1-\rho)}} = e^{\frac{2-\rho}{2(1-\rho)}\cdot\left(-\frac{a^2}{2}-\frac{b^2}{2}\right)} = e^{\left(1+\frac{\rho}{2(1-\rho)}\right)\cdot\left(-\frac{a^2}{2}-\frac{b^2}{2}\right)} =$$
$$\Pr_{x\in[N]^k}[x \in B]^{1+\frac{\rho}{2(1-\rho)}} \Pr_{x\in[N]^k}[x \in A]^{1+\frac{\rho}{2(1-\rho)}} .$$

Together we get

$$\Pr_{x,y}[x \in A, y \in B] \geq \Pr_{x\in[N]^k}[x \in A]^{1+\frac{\rho}{2(1-\rho)}} \Pr_{x\in[N]^k}[x \in B]^{1+\frac{3\rho}{2(1-\rho)}} .$$

□

3 Restricted global structure

In this section we prove the restricted global structure theorem, Theorem 1.6, which we restate formally below as Theorem 3.9. Let $f : [N]^k \to [M]^k$ be a function that passes Test 2 with probability ϵ, i. e.,

$$\mathrm{agr}^{\vee}_{k/10}(f) = \epsilon \geq e^{-c\alpha k}.$$

We show that such f already has some direct product structure, namely that there are restrictions of the domain $[N]^k$ such that f is close to a direct product function on each of the restricted parts.

Definition 3.1 (Restriction)**.** A *restriction* is a triple $\tau = (A, r, \gamma)$, for $A \subset [k]$, $|A| = k/10$, $r \in [N]^A$ and $\gamma \in [M]^A$.

Definition 3.2 (Consistent strings)**.** For every restriction $\tau = (A, r, \gamma)$, a string $w \in [N]^{\bar{A}}$ is consistent with τ if $f(r, w)_A = \gamma$. For every τ, let $\mathcal{V}_\tau$ be the set of consistent strings,

$$\mathcal{V}_\tau = \left\{ w \in [N]^{\bar{A}} \;\middle|\; f(r, w)_A = \gamma \right\} .$$

Definition 3.3 (Restricted function)**.** For each restriction $\tau = (A, r, \gamma)$, let $f_\tau : \mathcal{V}_\tau \to [M]^{\bar{A}}$ be the function

$$f_\tau(w) = f(r, w)_{\bar{A}} .$$

Definition 3.4 (Distribution over restrictions)**.** Let $\mathcal{D}$ be the following distribution over restrictions τ. Pick a uniform set $A \subset [k]$ of size $k/10$, pick a uniform $x \in [N]^k$ and output $\tau = (A, x_A, f(x)_A)$.

Note that the distribution $\mathcal{D}$ depends on the function f.

We define good restriction in an analogous way to the definitions of [13].

Definition 3.5 (Good restriction)**.** A restriction $\tau = (A, r, \gamma)$ is *good*, if $\Pr_{w\in[N]^{\bar{A}}}[w \in \mathcal{V}_\tau] \geq \epsilon/2$.

Definition 3.6 (α-DP restriction)**.** A restriction $\tau = (A, r, \gamma)$ is an *α-DP restriction* if it is good, and there exist functions $g_i : [N] \to [M]$ for each $i \in \bar{A}$ such that, denoting $g = (g_i)_{i \in \bar{A}}$,

$$\Pr_{w \in [N]^{\bar{A}}} \left[f_\tau(w) \overset{\alpha k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau \right] \leq \epsilon^2 .$$

Remark 3.7. The parameter α can be viewed as slack. It is the amount of disagreement we are willing to tolerate between two tuples, while still considering them in agreement.

We define a local averaging operator. Notice that the operator is not linear.

Definition 3.8 (Local averaging operator)**.** For every $\rho \in [0,1]$, let $\mathcal{P}_\rho$ be the following (non-linear) operator. For every subset $\mathcal{V}_\tau \subset [N]^{\bar{A}}$ and function $h : \mathcal{V}_\tau \to [M]^{\bar{A}}$, the operator takes h to the function $\mathcal{P}_\rho h : [N]^{\bar{A}} \to [M]^{\bar{A}}$ satisfies $\forall i \in \bar{A}, w \in [N]^{\bar{A}}$,

$$\mathcal{P}_\rho h(w)_i = \operatorname*{Plurality}_{v \underset{\rho,J}{\sim} w \text{ s.t. } i \in J} (h(v)_i) .$$

We define plurality over an empty set to be an arbitrary value.

In the proof of the restricted global structure we use the operator $\mathcal{P}_\rho$ with $\rho = 3/4$, which we denote by $\mathcal{P}$ to simplify notation. Clearly $3/4$ is an arbitrary constant, our proof works for any constant $\rho > 1/2$.

Our main theorem of this section asserts that (a) a non-negligible fraction of restrictions are good, and that (b) almost all good restrictions are DP restrictions.

Theorem 3.9 (Restricted Global Structure, formal)**.** *There exists a small constant $c > 0$, such that for every constant $\alpha > 0$ and large enough $k \in \mathbb{N}$ the following holds. For every function $f : [N]^k \to [M]^k$, if $agr^{\vee}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\alpha k}$, then*

$$\Pr_{\tau \sim \mathcal{D}} [\tau \text{ is good}] \geq \frac{\epsilon}{2}$$

and

$$\Pr_{\tau \sim \mathcal{D}} [\tau \text{ is an } \alpha\text{-DP restriction} \mid \tau \text{ is good}] \geq 1 - \epsilon^2 .$$

A similar theorem was proven in [8] under the name "local structure". Under the same assumptions [8] showed that f must be close to a product function for many pieces $\mathcal{V}_\tau$ of the domain. However, the closeness was considerably weaker: unlike in our definition of a *DP restriction*, in [8] even in the restricted part of the domain, $\mathcal{V}_\tau \subset [N]^k$, there could be a (small) constant fraction of the inputs on which f differs from the product function g. In contrast, we only allow an ϵ^2 fraction of disagreeing inputs, which is necessary for the ensuing global-structure argument.

Parameters. In the proof of the theorem we use several values for the slack parameter α, that are constant multiples of each other, which we denote by α_0, α_1 etc. These constant factors are not important, and the reader can treat all α_i as "similar to α".

We prove the theorem in Section 3.1, using lemmas that are proven on Section 3.2, Section 3.3 and Section 3.4.

3.1 Proof of Theorem 3.9

Let $f : [N]^k \to [M]^k$ be a function that passes Test 2 with probability ϵ. The test can also be written as: choose $\tau = (A, r, \gamma) \sim \mathcal{D}$ and a uniform $w \in [N]^{\bar{A}}$, and accept if $f(r, w)_A = \gamma$.

$$\epsilon = \Pr_{\tau\sim\mathcal{D}}\left[\text{test passes} \,\middle|\, \tau \text{ is chosen}\right] \le \Pr_{\tau\sim\mathcal{D}}\left[\tau \text{ is good}\right] \cdot 1 + \frac{\epsilon}{2},$$

because the success probability of the test on a restriction τ which is not good is at most $\epsilon/2$. Therefore,

$$\Pr_{\tau\sim\mathcal{D}}\left[\tau \text{ is good}\right] \ge \frac{\epsilon}{2}. \tag{3.1}$$

Fix $\alpha_0 = \frac{1}{1600}\alpha$.

Definition 3.10 (Excellent restriction)**.** A good restriction $\tau = (A, r, \gamma)$ is excellent if for $\rho = 3/4$ and for $\rho = 9/32$ the following holds. Choose $w \in [N]^{\bar{A}}$ uniformly and $v \underset{\rho,J}{\sim} w$ then,

$$\Pr_{w,v,J}\left[w, v \in \mathcal{V}_\tau \text{ and } f_\tau(w)_J \overset{\alpha_0 k}{\not\approx} f_\tau(v)_J\right] \le \mathrm{e}^{-\frac{\alpha_0 k}{40}} \overset{\Delta}{=} \mu\,. \tag{3.2}$$

Lemma 3.11. *A good $\tau \sim \mathcal{D}$ is excellent with probability at least $1 - \epsilon^2$.*

The proof of the lemma appears on Section 3.2.

To prove Theorem 3.9 it is enough to show that every excellent restriction is an *α-DP restriction*. A natural idea is to define a direct product function by taking the plurality of f_τ on $\mathcal{V}_\tau$, because the agreement of f_τ inside $\mathcal{V}_\tau$ is almost 1. However, it is difficult to prove that this function is close to f_τ because the set $\mathcal{V}_\tau$ is very sparse. Instead, we prove that $\mathcal{P} f_\tau$ is close to f_τ, and that $\mathcal{P} f_\tau$ is close to a direct product function. Fix $\alpha_1 = 10\alpha_0$.

Lemma 3.12. *For every excellent τ,*

$$\Pr_{w\in[N]^{\bar{A}}}\left[f_\tau(w) \overset{\alpha_1 k}{\not\approx} \mathcal{P} f_\tau(w) \,\middle|\, w \in \mathcal{V}_\tau\right] \le \epsilon^3\,.$$

The proof appears in Section 3.3 and relies on reverse hypercontractivity.

Fix $\alpha_2 = 1500\alpha_0$.

Lemma 3.13. *For every excellent restriction τ there exists a direct product function $g = (g_1, \ldots, g_{\bar{A}})$, $g_i : [N] \to [M]$ such that*

$$\Pr_{w\in[N]^{\bar{A}}}\left[\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w)\right] \le 3\epsilon^4\,.$$

The proof is in Section 3.4.

The two lemmas above, Lemma 3.12 and Lemma 3.13, imply the following claim.

Claim 3.14. *Every excellent restriction τ is an α-DP restriction.*

Proof. Fix an excellent restriction τ. Let $g = (g_1, \ldots, g_{\bar{A}})$, $g_i : [N] \to [M]$ be the direct product function promised from Lemma 3.13. For an excellent τ, $\Pr_w[w \in \mathcal{V}_\tau] \geq \frac{\epsilon}{2}$. Therefore, the probability of $\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w)$ is small even when conditioning on $w \in \mathcal{V}_\tau$. By Lemma 3.13

$$\begin{aligned}
3\epsilon^4 \geq & \Pr_{w\in[N]^{\bar{A}}}\left[\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w)\right] \\
\geq & \Pr_{w\in[N]^{\bar{A}}}[w \in \mathcal{V}_\tau] \Pr_{w\in[N]^{\bar{A}}}\left[\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right] \\
\geq & \frac{\epsilon}{2} \Pr_{w\in[N]^{\bar{A}}}\left[\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right]. \qquad (3.3)
\end{aligned}$$

By the triangle inequality,

$$\begin{aligned}
\Pr_w\left[f_\tau(w) \overset{(\alpha_1+\alpha_2)k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right] \leq & \Pr_w\left[f_\tau(w) \overset{\alpha_1 k}{\not\approx} \mathcal{P} f_\tau(w) \,\middle|\, w \in \mathcal{V}_\tau\right] \\
& + \Pr_w\left[\mathcal{P} f_\tau(w) \overset{\alpha_2 k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right] \\
\leq & \epsilon^3 + 6\epsilon^3 < \epsilon^2. \qquad \text{(by (3.3) and Lemma 3.12)}
\end{aligned}$$

By definition, $f_\tau(w) = f(x_A, w)_{\bar{A}}$, so from the above equation

$$\Pr_w\left[f(x_A, w)_{\bar{A}} \overset{(\alpha_1+\alpha_2)k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right] = \Pr_w\left[f_\tau(w) \overset{(\alpha_1+\alpha_2)k}{\not\approx} g(w) \,\middle|\, w \in \mathcal{V}_\tau\right] < \epsilon^2.$$

Since $\alpha_1 + \alpha_2 < \alpha$ we are done. □

In the proof we see that a random restriction $\tau \sim \mathcal{D}$ is good with probability $\epsilon/2$, and that a good restriction is excellent with probability $1 - \epsilon^2$. From the claim above, an excellent restriction τ is an α-DP restriction, which finishes the proof.

□

3.2 Good restrictions are excellent with high probability

In this section we prove Lemma 3.11, which states that a good restriction is excellent with high probability. We start by showing that when averaging over τ, f_τ is consistent in $\mathcal{V}_\tau$.

Claim 3.15. *For every $\rho \in (0,1)$, let $\tau \sim \mathcal{D}$, $w \in [N]^{\bar{A}}$ and $v \underset{\rho,J}{\sim} w$, then*

$$\Pr_{\tau,w,v,J}\left[w, v \in \mathcal{V}_\tau,\ f_\tau(w)_J \overset{\alpha_0 k}{\not\approx} f_\tau(v)_J\right] \leq \mathrm{e}^{-\frac{\alpha_0 k}{20}}.$$

Proof. Denote $\tau = (r, A, \gamma)$. Let $E_1(r, A, w, v, J, \gamma)$ be the event that we are interested in bounding its probability, namely, the event

$$f(r,w)_A = f(r,v)_A = \gamma \quad \text{and} \quad f(r,w)_J \overset{\alpha_0 k}{\not\approx} f(r,v)_J \,.$$

When r, w, v, A, J are all random, the probability of a uniform A, J to be such that $f(r,w), f(r,v)$ are equal on A but are far on J is very small.

Let $E_2(A, r, v, w, J) \subseteq E_1(A, r, v, w, J, \gamma)$ be the event that $f(r,w)_A = f(r,v)_A$ and $f(r,w)_J \overset{\alpha_0 k}{\not\approx} f(r,v)_J$.

Fix $\rho \in (0,1)$. We start by bounding the probability of E_2 under the distribution $\tau \sim \mathcal{D}$, $w \in [N]^{\bar{A}}$ uniformly and $v \underset{\rho,J}{\sim} w$. Writing the distribution explicitly:

1. Pick $A \subset [k]$ of size $k/10$.
2. Pick $x \in [N]^k$, set $r = x_A$ and $\gamma = f(x)_A$.
3. Pick $J \subset \bar{A}$ of size $\mathcal{B}(9k/10, \rho)$ (binomial random variable).
4. Pick uniform $w, v \in [N]^{\bar{A}}$ such that $w_J = v_J$.

Notice that E_2 is independent of γ, so it does not matter how γ is chosen. We can define an equivalent process for producing the same distribution (without γ):

1. Pick a set $A' \subset [k]$ of size $k/10 + \mathcal{B}(9k/10, \rho)$.
2. Pick $y, z \in [N]^k$ such that $y_{A'} = z_{A'}$.
3. Pick $A \subseteq A'$ of size $k/10$.
4. Set $r = y_A$, $w = y_{\bar{A}}$ and $v = z_{\bar{A}}$.

Let

$$D = \{i \in A' \mid \ f(y)_i \neq f(z)_i\} \,.$$

E_2 occurs only if $A \cap D = \emptyset$ yet $|D| \geq \alpha_0 k$.

As the second random process allows us to see, A is a uniform subset of A'. Let $i_1, \ldots i_{|A|}$ be an arbitrary order over the elements of A. We can think of A as being chosen incrementally, each i_j is chosen randomly from $A' \setminus \{i_1, \ldots, i_{j-1}\}$.

$$\begin{aligned}
\Pr_A[A \cap D = \emptyset] &= \Pr\left[i_1, \ldots, i_{|A|} \notin D\right] \qquad (3.4)\\
&= \Pr[i_1 \notin D] \cdot \Pr[i_2 \notin D \mid i_1 \notin D] \cdots \Pr\left[i_{|A|} \notin D \,\middle|\, i_1, \ldots, i_{|A|-1} \notin D\right]\\
&\leq (1-\alpha_0)^{|A|} \leq \mathrm{e}^{-\alpha_0 |A|}
\end{aligned}$$

where we use the fact that for every i_j, $\Pr\left[i_j \notin D \,\middle|\, i_1 \ldots i_{j-1} \notin D\right] \leq 1 - \alpha_0$, since $|D| \geq \alpha_0 k$.

The bound in (3.4) holds for each A' such that $f(y)_{A'} \overset{\alpha_0 k}{\not\approx} f(z)_{A'}$, therefore also on average,

$$\Pr[E_2] = \Pr_{y,z,A',A}\left[f(y)_{A'} \overset{\alpha_0 k}{\not\approx} f(z)_{A'} \text{ and } f(y)_A = f(z)_A\right] \leq \mathrm{e}^{-\alpha_0|A|}.$$

We conclude that $\Pr[E_1] \leq \Pr[E_2] \leq \mathrm{e}^{-\alpha_0|A|} \leq \mathrm{e}^{-\frac{\alpha_0 k}{20}}$. ☐

Proof of Lemma 3.11. For every restriction τ and $w, v \in [N]^{\bar{A}}$ let $E(\tau, w, v, J)$ be the event that $w, v \in \mathcal{V}_\tau$ and $f_\tau(w)_J \overset{\alpha_0 k}{\not\approx} f_\tau(v)_J$.

For every τ that is good but not excellent, there is $\rho \in \{3/4, 9/32\}$ such that

$$\Pr_{w, v \underset{\rho, J}{\sim} w}[E(\tau, w, v, J)] > \mu\,.$$

In this case we say that τ is bad for ρ.

Assume for a contradiction that $\Pr_{\tau\sim\mathcal{D}}\left[\tau \text{ is good but not excellent}\right] > \epsilon^3/2$. These fail to be excellent because of at least one of the choices of ρ, so either for $\rho = 3/4$ or for $\rho = 9/32$,

$$\Pr_{\tau\sim\mathcal{D}}[\tau \text{ is bad for } \rho] \geq \frac{\epsilon^3}{4}\,.$$

For this ρ,

$$\Pr_{\tau\sim\mathcal{D}, w, v \underset{\rho, J}{\sim} w}[E(\tau, w, v, J)] \geq \Pr_{\tau\sim\mathcal{D}}[\tau \text{ is bad for } \rho] \Pr_{w, v \underset{\rho, J}{\sim} w}[E(\tau, w, v, J) \mid \tau \text{ is bad for } \rho] \geq \frac{\epsilon^3}{4}\mu\,.$$

This contradicts Claim 3.15, because $\epsilon^3/4 \cdot \mu \geq \mathrm{e}^{-\frac{\alpha k}{20}}$, as $\alpha_0 = 1600\alpha$. Therefore, we conclude that $\Pr_{\tau\sim\mathcal{D}}\left[\tau \text{ is good but not excellent}\right] \leq \epsilon^3/2$.

Finally, since $\tau \sim \mathcal{D}$ is good with probability at least $\epsilon/2$, by averaging a good $\tau \sim \mathcal{D}$ is excellent with probability at least $1 - \epsilon^2$. ☐

3.3 Local averaging operator

In this section we prove Lemma 3.12, which states that for every excellent τ, $\mathcal{P} f_\tau$ approximates f_τ.

Fix an excellent τ, and let $L \subset [N]^{\bar{A}}$ be the set of "lonely" strings in $[N]^{\bar{A}}$, those that have a sparse neighborhood in $\mathcal{V}_\tau$. We fix te sparsity parameter $\eta = \epsilon^{50}$, and define

$$L = \left\{ w \in [N]^{\bar{A}} \,\middle|\, \Pr_{v \underset{3/4, J}{\sim} w}[v \in \mathcal{V}_\tau] \leq \eta \right\}.$$

There is a trade off between the "sparsity" parameter η and the bound on the size of L. The sparsity parameter is chosen to make sure that $\Pr_{w\in[N]^{\bar{A}}}[w \in L] < \epsilon^{10}$. These powers of ϵ

are derived from our arbitrary choice of defining $\mathcal{P}f_\tau$ as the plurality of a 3/4-correlated neighborhood.

In addition, the success probability of the test ϵ should be large enough to promise that μ/η^2 is small.

Claim 3.16. $\Pr_{w\in[N]^{\bar{A}}}[w \in L] < \epsilon^{10}$

Proof. We prove the claim by using Corollary 2.9 on the sets $L,\mathcal{V}_\tau$. Denote by $p_L = \Pr_{w\in[N]^{\bar{A}}}[w \in L]$ and $p = \Pr_{w\in[N]^{\bar{A}}}[w \in \mathcal{V}_\tau]$. Since τ is excellent, $p \geq \epsilon/2$.

Assume for a contradiction that $|L| > |\mathcal{V}_\tau|$, i. e., $p_L \geq p$. By Corollary 2.9

$$\Pr_{w\in[N]^{\bar{A}},v\underset{3/4}{\sim}w}[w \in L, v \in \mathcal{V}_\tau] \geq p_L^{\frac{5}{2}} p^{\frac{11}{2}} \geq \left(\frac{\epsilon}{2}\right)^{\frac{5}{2}} \left(\frac{\epsilon}{2}\right)^{\frac{11}{2}} .$$

By the definition of L

$$\Pr_{w\in[N]^{\bar{A}},v\underset{3/4}{\sim}w}[w \in L, v \in \mathcal{V}_\tau] \leq p_L\epsilon^{50} , \tag{3.5}$$

and we reach a contraction.

Therefore, $p_L < p$. By Corollary 2.9,

$$\Pr_{w\in[N]^{\bar{A}},v\underset{3/4}{\sim}w}[w \in L, v \in \mathcal{V}_\tau] \geq p_L^{\frac{11}{2}} p^{\frac{5}{2}} \geq \left(\frac{\epsilon}{2}\right)^{\frac{5}{2}} p_L^{\frac{11}{2}} .$$

Equation (3.5) still holds, so combining the two bounds on $\Pr_{w\in[N]^{\bar{A}},v\underset{3/4}{\sim}w}[w \in L, v \in \mathcal{V}_\tau]$,

$$\left(\frac{\epsilon}{2}\right)^{\frac{5}{2}} p_L^{\frac{11}{2}} \leq p_L\epsilon^{50},$$

That is $p_L \leq 2\epsilon^{\frac{9}{2}(50-\frac{5}{2})} < \epsilon^{10}$. ◻

The local averaging operator $\mathcal{P}f_\tau$ takes the plurality vote over $v \underset{3/4,J}{\sim} w$, conditioning on $v \in \mathcal{V}_\tau$. If we pick $v \underset{3/4,J}{\sim} w$ without conditioning on $v \in \mathcal{V}_\tau$, then each i is in J with probability $3/4$. For $w \notin L$, taking $v \underset{3/4,J}{\sim} w$ conditioning on $v \in \mathcal{V}_\tau$, we get that for most i, $i \in J$ with probability that is close to $3/4$.

Claim 3.17. *For every $w \notin L$ and all except $\alpha_0 k$ of the coordinates $i \in \bar{A}$,*

$$\Pr_{v\underset{3/4,J}{\sim}w}[J \ni i | v \in \mathcal{V}_\tau] \geq \frac{3}{4} - \frac{1}{10} .$$

Proof. Fix $w \notin L$. Let D be

$$D = \left\{ i \in [k] \;\middle|\; \Pr_{v \underset{3/4,J}{\sim} w}[J \ni i | v \in \mathcal{V}_\tau] < \frac{3}{4} - \frac{1}{10} \right\} .$$

Assume for a contradiction that the claim does not hold, i. e., $|D| > \alpha_0 k$.

By the Chernoff bound (Fact 2.3), $\Pr_{v \underset{3/4,J}{\sim} w}\left[|D \cap J| \leq \left(\frac{3}{4} - \frac{1}{20}\right)|D|\right] \leq e^{-\frac{\alpha_0 k}{900}}$. For $w \notin L$, conditioning on $v \in \mathcal{V}_\tau$ increases the above probability by a factor of at most $1/\eta$. That is,

$$\Pr_{v \underset{3/4,J}{\sim} w}\left[|D \cap J| \leq \left(\frac{3}{4} - \frac{1}{20}\right)|D| \;\middle|\; v \in \mathcal{V}_\tau\right] \leq \frac{1}{\eta} e^{-\frac{\alpha_0 k}{900}} .$$

The constant c in is chosen to be small enough to promise that $\frac{1}{\eta} e^{-\frac{\alpha_0 k}{900}} = \epsilon^{-50} \cdot e^{-\frac{\alpha_0 k}{900}} \leq \epsilon$. That is, $|J \cap D| \geq \left(\frac{3}{4} - \frac{1}{20}\right)|D|$ almost always.

If J is such that $|J \cap D| \geq \left(\frac{3}{4} - \frac{1}{20}\right)|D|$, then a random $i \in D$ has probability of at least $\left(\frac{3}{4} - \frac{1}{20}\right)$ to be in J. Therefore,

$$\begin{aligned}\Pr_{v \underset{3/4,J}{\sim} w, i \in D}[i \in J \mid v \in \mathcal{V}] \geq & \Pr_{v \underset{3/4,J}{\sim} w}\left[|D \cap J| \leq \left(\frac{3}{4} - \frac{1}{20}\right)|D| \;\middle|\; v \in \mathcal{V}_\tau\right]\left(\frac{3}{4} - \frac{1}{20}\right) \\ \geq & (1-\epsilon)\left(\frac{3}{4} - \frac{1}{20}\right) > \frac{3}{4} - \frac{1}{10} ,\end{aligned}$$

which is a contradiction to the definition of D. □

Proof of Lemma 3.12. Fix an excellent restriction τ. Recall that for an excellent restriction τ, $\Pr_{w \in [N]^{\bar{A}}}[w \in \mathcal{V}_\tau] \geq \epsilon/2$ and

$$\Pr_{w, v \underset{3/4,J}{\sim} w}\left[w, v \in \mathcal{V}_\tau \text{ and } h(w)_J \overset{\alpha_0 k}{\not\approx} h(v)_J\right] \leq \mu . \tag{3.6}$$

Let B be the set of "bad" inputs,

$$B = \left\{ w \in \mathcal{V}_\tau \;\middle|\; \Pr_{v \underset{3/4,J}{\sim} w}\left[v \in \mathcal{V}_\tau \text{ and } f_\tau(v)_J \overset{\alpha k}{\not\approx} f_\tau(w)_J\right] \geq \frac{\eta}{10} \right\} .$$

By averaging on equation (3.6), $\Pr_{w \in \mathcal{V}_\tau}[w \in B] \leq \frac{10\mu}{\eta} \leq \frac{1}{2}\epsilon^3$.

Recall L is the set of lonely inputs. By Claim 3.16, $\Pr_{w \in [N]^k}[w \in L] \leq \epsilon^{10}$. Since τ is excellent,

$$\Pr_{w \in \mathcal{V}_\tau}[w \in L] \leq \frac{\epsilon^{10}}{\frac{\epsilon}{2}} < \frac{\epsilon^3}{2} .$$

Together, the probability that a random $w \in \mathcal{V}_\tau$ is in B or in L is smaller than ϵ^3. Therefore, to prove the lemma it is enough to prove that for every $w \in \mathcal{V}_\tau \setminus \{B \cup L\}$, $f_\tau(w) \overset{\alpha_1 k}{\approx} \mathcal{P} f_\tau(w)$.

Fix $w \in \mathcal{V}_\tau \setminus (B \cup L)$. Denote by D the set of coordinates on which $f_\tau(w), \mathcal{P} f_\tau(w)$ differ,

$$D = \left\{ i \in \bar{A} \,\middle|\, f_\tau(w)_i \neq \mathcal{P} f_\tau(w)_i \right\} .$$

Assume for a contradiction that $|D| > \alpha_1 k$. For $v \in [N]^{\bar{A}}$, $J \subset \bar{A}$ and $i \in \bar{A}$, let $E(v, J, i)$ be the event

$$i \in J \quad \text{and} \quad f_\tau(w)_i \neq f_\tau(v)_i .$$

We reach a contradiction by upper bounding and lower bounding the probability of E under the distribution $i \in D$ and $v \underset{3/4,J}{\sim} w$, given that $v \in \mathcal{V}_\tau$.

Lower bound. The function $\mathcal{P} f_\tau$ takes the most common value $f_\tau(v)_i$ for $v \underset{3/4,J}{\sim} w$. By definition, the set D contains all of the coordinates i such that $f_\tau(w)_i$ is *not* the most probable value $f_\tau(v)_i$, when $v \underset{3/4,J}{\sim} w$. Therefore for every $i \in D$,

$$\Pr_{v \underset{3/4,J}{\sim} w} [f_\tau(w)_i \neq f_\tau(v)_i \mid i \in J, v \in \mathcal{V}_\tau] \geq \frac{1}{2}. \tag{3.7}$$

To lower bound the probability of E we need to lower bound the probability of $i \in J$. When picking $v \underset{3/4,J}{\sim} w$, each i is in J with probability $3/4$. From Claim 3.17, for all except $\alpha_0 k$ coordinates, the probability of i to be in J is not much different, it is at least $3/4 - 1/10$. Let $D' \subset [k]$ be the set of coordinates in which $\Pr[i \in J]$ is lower than $3/4 - \frac{1}{10}$, then $|D'| \leq \alpha_0 k < 1/10|D|$. Together we get that

$$\Pr_{i \in D, v \underset{3/4,J}{\sim} w} [i \in J \mid v \in \mathcal{V}_\tau] \geq \Pr_{i \in D}[i \notin D'] \left(\frac{3}{4} - \frac{1}{10} \right) > \frac{1}{2} .$$

From the two equations above,

$$\begin{aligned}
\Pr_{v \underset{3/4,J}{\sim} w, i \in D} [E(v, J, i) \mid v \in \mathcal{V}_\tau] &= \Pr_{v \underset{3/4,J}{\sim} w, i \in D} [i \in J \text{ and } f_\tau(w)_i \neq f_\tau(v)_i \mid v \in \mathcal{V}_\tau] \\
&= \Pr_{v \underset{3/4,J}{\sim} w, i \in D} [i \in J \mid v \in \mathcal{V}_\tau] \Pr_{v \underset{3/4,J}{\sim} w} [f_\tau(w)_i \neq f_\tau(v)_i \mid i \in J, v \in \mathcal{V}_\tau] \\
&> \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} .
\end{aligned}$$

Upper Bound. For $w \in \mathcal{V}_\tau \setminus (B \cup L)$, the following two equations hold: $\Pr_{v \underset{3/4,J}{\sim} w} [v \in \mathcal{V}_\tau] \geq \eta$ and $\Pr_{v \underset{3/4,J}{\sim} w} \left[v \in \mathcal{V}_\tau \text{ and } f_\tau(v)_J \overset{\alpha_0 k}{\not\approx} f_\tau(w)_J \right] \leq \frac{\eta}{10}$. Combining the two equations,

$$\Pr_{v \underset{3/4,J}{\sim} w} \left[f_\tau(v)_J \overset{\alpha_0 k}{\not\approx} f_\tau(w)_J \,\middle|\, v \in \mathcal{V}_\tau \right] \leq \frac{1}{10} . \tag{3.8}$$

Suppose $f_\tau(v)_J \overset{\alpha_0 k}{\approx} f_\tau(w)_J$, our assumption is that $|D| > \alpha_1 k = 10\alpha_0 k$, so a uniform $i \in D$ is in the $\alpha_0 k$ coordinates in which $f_\tau(v)_J, f_\tau(w)_J$ differ with probability at most $\frac{1}{10}$. This lets us bound the probability of E:

$$\begin{aligned}\Pr_{v \underset{3/4,J}{\sim} w, i\in D}[E(v,J,i) \mid v \in \mathcal{V}_\tau] \leq & \Pr_{v \underset{3/4,J}{\sim} w}\left[f_\tau(v)_J \overset{\alpha_0 k}{\not\approx} f_\tau(w)_J \,\middle|\, v \in \mathcal{V}_\tau \right] \\ & + \Pr_{v \underset{3/4,J}{\sim} w, i\in D}\left[f_\tau(v)_J \overset{\alpha_0 k}{\approx} f_\tau(w)_J \text{ and } i \in J \text{ and } f_\tau(w)_i \neq f_\tau(v)_i \,\middle|\, v \in \mathcal{V}_\tau \right] \\ \leq & \frac{1}{10} + \frac{1}{10} < \frac{1}{4}.\end{aligned}$$

The upper and lower bounds contradict each other so for all $w \in \mathcal{V}_\tau \setminus (B \cup L)$, $|D| \leq \alpha_1 k$ and $f_\tau(w) \overset{\alpha_1 k}{\approx} \mathcal{P} f_\tau(w)$. Since $\Pr_w[w \in B \cup L] \leq \epsilon^3$ we finish the proof. □

3.4 Direct product function

In this section we prove that $\mathcal{P} f_\tau$ is close to a direct product function, proving Lemma 3.13.

We start by defining the candidate direct product function, which is the plurality vote of $\mathcal{P} f_\tau$.

Definition 3.18. For every excellent $\tau = (A, r, \gamma)$, let $g_\tau : [N]^{\bar{A}} \to [M]^{\bar{A}}$ be the following function: for every $i \notin A$ and $b \in [N]$,

$$g_{\tau,i}(b) = \operatorname*{Plurality}_{w \in [N]^{\bar{A}} \text{ s.t. } w_i = b} \{\mathcal{P} f_\tau(w)_i\},$$

ties are broken arbitrarily.

Set $\alpha_3 = 20\alpha_0$.

Claim 3.19. *For every excellent τ,*

$$\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}\left[\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\approx} \mathcal{P} f_\tau(v)_J \right] \geq 1 - 3\epsilon^{10}.$$

Proof. The proof is similar to the proof of Lemma 3.12. Its main idea is that if $\mathcal{P} f_\tau(w)$ and $\mathcal{P} f_\tau(v)$ disagree on a lot of coordinates, then a large fraction of their 3/4-correlated neighborhoods also disagree on a lot of coordinates. This can only happen for very few inputs w, v, otherwise we contradict the fact that τ is excellent.

Fix an excellent τ. Let C be a set of "bad" triplets,

$$C = \left\{ (w, v, J) \,\middle|\, w_J = v_J \text{ and } \Pr_{w', J', v', J''}\left[w', v' \in \mathcal{V}_\tau \text{ and } f_\tau(w')_{\tilde{J}} \overset{\alpha_0 k}{\not\approx} f_\tau(v')_{\tilde{J}} \right] \geq \frac{\eta^2}{10} \right\},$$

where $w' \underset{3/4,J'}{\sim} w, v' \underset{3/4,J''}{\sim} v$ and $\tilde{J} = J \cap J' \cap J''$.

If (w, v, J) are distributed such that w is uniform in $[N]^{\bar{A}}$ and $v \underset{1/2,J}{\sim} w$, then the marginal distribution over w' is uniform, and $v' \underset{9/32,\tilde{J}}{\sim} w'$. This is because for every i independently, the probability of i to be in $\tilde{J} = J \cap J' \cap J''$ is $(3/4)^2 \cdot (1/2) = 9/32$.

Since τ is excellent, $\Pr_{w' \in [N]^{\bar{A}}, v' \underset{9/32,\tilde{J}}{\sim} w'}\left[w', v' \in \mathcal{V}_\tau, f_\tau(w')_{\tilde{J}} \overset{\alpha_0 k}{\not\approx} f_\tau(v')_{\tilde{J}}\right] \leq \mu$. Therefore by averaging, $\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}[(w, v, J) \in C] \leq \frac{10\mu}{\eta^2} < \epsilon^{10}$.

Recall that L is the set of "lonely" inputs, those with sparse neighborhood. From Claim 3.16, $\Pr_{w \in [N]^{\bar{A}}}[w \in L] < \epsilon^{10}$.

We prove next that for every w, v, J such that $(w, v, J) \notin C$ and $w, v \notin L$, $\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\approx} \mathcal{P} f_\tau(v)_J$. This proves the claim since by a union bound argument,

$$\begin{aligned} &\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}[(w, v, J) \in C \text{ or } w \in L \text{ or } v \in L] \\ &\leq 2 \Pr_{w \in [N]^{\bar{A}}}[w \in L] + \Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} v}[(w, v, J) \in C] \leq 3\epsilon^{10}. \end{aligned}$$

Fix w, v, J such that $w_J = v_J$, $w, v \notin L$ and $(w, v, J) \notin C$. Let $D \subseteq J$ be the set

$$D = \{i \in J \mid \mathcal{P} f_\tau(w)_i \neq \mathcal{P} f_\tau(v)_i\}.$$

Assume for a contradiction that $|D| \geq \alpha_3 k$.

For every $J', J'' \subset \bar{A}$, $w', v' \in \mathcal{V}$ and $i \in \bar{A}$, let $E(J', J'', w', v', i)$ be the following event:

$$f_\tau(w')_i \neq f_\tau(v')_i \quad \text{and} \quad i \in J' \cap J''.$$

We reach a contradiction by giving an upper bound and a lower bound of this event, under the distribution $i \in D$, $w' \underset{3/4,J'}{\sim} w$ and $v' \underset{3/4,J''}{\sim} v$, given that $w', v' \in \mathcal{V}_\tau$.

Lower Bound. For every $i \in D$, $\mathcal{P} f_\tau(w)_i \neq \mathcal{P} f_\tau(v)_i$, i. e., the most frequent value $f_\tau(w')_i$ for $w' \underset{3/4}{\sim} w$ is *different* from the most frequent value $f_\tau(v')_i$ for $v' \underset{3/4}{\sim} v$. Therefore, for every $i \in D$,

$$\Pr_{w', J', v', J''}[f_\tau(w')_i \neq f_\tau(v')_i \mid i \in J' \cap J'', w', v' \in \mathcal{V}_\tau] \geq \frac{1}{2}, \tag{3.9}$$

where $w' \underset{3/4,J'}{\sim} w$ and $v' \underset{3/4,J''}{\sim} v$.

To finish the lower bound, we need to lower bound the probability of $i \in J' \cap J''$, for $w' \underset{3/4,J'}{\sim} w$ and $v' \underset{3/4,J''}{\sim} v$ given that $w', v' \in \mathcal{V}_\tau$. Without conditioning on $w', v' \in \mathcal{V}_\tau$, each i is in $J' \cap J''$ with probability $(3/4)^2$. We show that the probability is not much lower even when conditioning on $w', v' \in \mathcal{V}_\tau$. From Claim 3.17, for all except $\alpha_0 k$ of the coordinates $i \in \bar{A}$, the probability of i to be in J' is at least $3/4 - 1/10$. Let $D_{J'}$ be the set of coordinates such that $\Pr[i \in J'] \leq \frac{3}{4} - \frac{1}{10}$ and let $D_{J''}$ be the equivalent set for J''. Then $|D \setminus (D_{J'} \cup D_{J''})| \geq 20\alpha_0 k - \alpha_0 k - \alpha_0 k$.

$$\Pr_{i \in D, w' \underset{3/4,J'}{\sim} w, v' \underset{3/4,J''}{\sim} v} [i \in J' \cap J''] \geq \Pr_{i \in D}\left[i \notin D_{J'} \cap D_{J''}\right]\left(\frac{3}{4} - \frac{1}{10}\right)^2 > \frac{1}{3}.$$

Combining the two equations, we lower bound the probability of E, under the distribution $i \in D, w' \underset{3/4,J'}{\sim} w$ and $v' \underset{3/4,J''}{\sim} v$,

$$\begin{aligned}\Pr_{i,w',J',v',J''}[E(J',J'',w',v',i) \mid w',v' \in \mathcal{V}_\tau] &= \Pr_{i,w',J',v',J''}[i \in J' \cap J'' \mid w',v' \in \mathcal{V}_\tau] \\ &\quad \cdot \Pr_{i,w',J',v',J''}[f_\tau(w')_i \neq f_\tau(v')_i \mid w',v' \in \mathcal{V}_\tau, i \in J' \cap J''] \\ &\geq \frac{1}{3} \cdot \frac{1}{2} \geq \frac{1}{6}.\end{aligned}$$

Upper Bound. The fixed w, v, J satisfy $\Pr_{v' \underset{3/4,J''}{\sim} v}[v' \in \mathcal{V}_\tau] \geq \eta$ and

$$\Pr_{w' \underset{3/4,J'}{\sim} w, v' \underset{3/4,J''}{\sim} v}\left[w', v' \in \mathcal{V}_\tau \text{ and } f_\tau(w')_{\tilde{J}} \overset{\alpha_0 k}{\not\approx} f_\tau(v')_{\tilde{J}}\right] \leq \frac{\eta^2}{10},$$

where $\tilde{J} = J \cap J' \cap J''$.

Therefore,

$$\Pr_{w' \underset{3/4,J'}{\sim} w, v' \underset{3/4,J''}{\sim} v}\left[f_\tau(w')_{\tilde{J}} \overset{\alpha_0 k}{\not\approx} f_\tau(v')_{\tilde{J}} \,\middle|\, w', v' \in \mathcal{V}_\tau\right] \leq \frac{1}{10}.$$

If $f_\tau(w')_{\tilde{J}} \overset{\alpha_0 k}{\approx} f_\tau(v')_{\tilde{J}}$, then a uniform $i \in D$ has probability of at most $\frac{\alpha_0 k}{|D|} \leq \frac{\alpha_0 k}{20\alpha_0 k} \leq 1/20$ to be in a coordinate in which $f_\tau(w'), f_\tau(v')$ differ. Therefore,

$$\Pr_{i \in D, w' \underset{3/4,J'}{\sim} w, v' \underset{3/4,J''}{\sim} v}[E(J',J'',w',v',i) \mid w',v' \in \mathcal{V}_\tau] \leq \frac{1}{10} + \frac{1}{20} < \frac{1}{6},$$

which contradicts the lower bound. This completes the proof of Claim 3.19. ☐

Claim 3.20. *For every excellent* τ,

$$\Pr_{i \in \bar{A}, w, v \in [N]^{\bar{A}}}[\mathcal{P} f_\tau(w)_i = \mathcal{P} f_\tau(v)_i \mid w_i = v_i] \geq 1 - 10\alpha_3.$$

Proof. Fix an excellent τ. Claim 3.19 proves the agreement of $\mathcal{P} f_\tau$ on *correlated* inputs. In this claim, we prove its agreement on *uncorrelated* inputs.

Let $\mathcal{D}'$ be the distribution of picking $u, v, w \in [N]^{\bar{A}}$ and $i \in \bar{A}$ defined as follows.

1. Pick a uniform $i \in \bar{A}$.
2. Pick uniform $w, v \in [N]^{\bar{A}}$ such that $w_i = v_i$.
3. For every $j \neq i$, insert j into J with probability $\frac{1}{2}$ independently.
4. For every $j \in \bar{A}$ set $u_j = \begin{cases} w_j & j \in J \\ v_j & \text{else} \end{cases}$.

The distribution $\mathcal{D}'$ produces w, v that are *uncorrelated* except that $w_i = v_i$. The pair w, u are nearly 1/2-correlated (the only difference is the coordinate i in which $u_i = w_i$ with probability 1). We prove the claim by applying Claim 3.19 on the correlated pairs u, w and u, v, and get that with high probability $\mathcal{P}f_\tau(u)_i = \mathcal{P}f_\tau(w)_i$ and $\mathcal{P}f_\tau(u)_i = \mathcal{P}f_\tau(v)_i$, deducing that $\mathcal{P}f_\tau(w)_i = \mathcal{P}f_\tau(v)_i$.

The marginal distribution on w, u in $\mathcal{D}'$ is very close to 1/2-correlated. The only difference between the marginal distribution of $w, u, J, i \sim \mathcal{D}'$ and $u \underset{1/2}{\sim} w$ is that in $\mathcal{D}'$, $w_i = u_i$ with probability 1 and not $1/2$. In Claim 3.23 we prove that the probability of any event on $u \underset{1/2,J}{\sim} w$ can increase by at most a factor of 2, for $u, w, J \cup \{i\}$ produced by $\mathcal{D}'$.

Therefore, we can use Claim 3.19 on $w, u, J, i \sim \mathcal{D}'$ and pay a factor of 2,

$$\Pr_{w,u,J,i\sim\mathcal{D}'}\left[\mathcal{P}f_\tau(w)_{J\cup\{i\}} \overset{\alpha_3 k}{\not\approx} \mathcal{P}f_\tau(u)_{J\cup\{i\}}\right] \leq 2\cdot 3\epsilon^{10}. \tag{3.10}$$

The same holds also for $v, u, \bar{A}\setminus J, i \sim \mathcal{D}'$.

For $w, u, J, i \sim \mathcal{D}'$,the coordinate i is a random coordinate in $J \cup \{i\}$. Therefore, if $\mathcal{P}f_\tau(w)_{J\cup\{i\}} \overset{\alpha_3 k}{\approx} \mathcal{P}f_\tau(u)_{J\cup\{i\}}$, then the probability of i be be such that $\mathcal{P}f_\tau(w)_i \neq \mathcal{P}f_\tau(u)_i$ is at most $\frac{\alpha_3 k}{|J\cup\{i\}|}$. Each $j \neq i$ is in J with probability $1/2$ independently, which lets us bound the size of J by the Chernoff bound, $\Pr_J[|J| < k/4] \leq \mathrm{e}^{\frac{1}{3\cdot 5^2}\frac{1}{2}|\bar{A}|} < \epsilon$. If $|J| > k/4$, then the probability of a random $i \in J$ to fall into the $\alpha_3 k$ disagreeing coordinates is at most $4\alpha_3 k$. Therefore,

$$\begin{aligned}\Pr_{w,u,J,i\sim\mathcal{D}'}[\mathcal{P}f_\tau(w)_i \neq \mathcal{P}f_\tau(u)_i] \leq \Pr_{w,u,J,i\sim\mathcal{D}'}&\left[|J| < \frac{k}{4} \text{ or } \mathcal{P}f_\tau(w)_{J\cup\{i\}} \overset{\alpha_3 k}{\not\approx} \mathcal{P}f_\tau(u)_{\cup\{i\}}\right] \\ &+ 4\alpha_3 k \leq 6\epsilon^{10} + \epsilon + 4\alpha_3 k < 5\alpha_3 k.\end{aligned} \tag{3.11}$$

The same holds also for $v, u, \bar{A}\setminus J, i$.

Therefore, from equation (3.11) on w, u, J, i and $v, u, \bar{A}\setminus J, i$,

$$\begin{aligned}\Pr_{\substack{i\in\bar{A}\\ w,v\in[N]^{\bar{A}}}}[\mathcal{P}f_\tau(w)_i = \mathcal{P}f_\tau(v)_i \mid w_i = v_i] \geq& \Pr_{i,w,v,u\sim\mathcal{D}'}[\mathcal{P}f_\tau(w)_i = \mathcal{P}f_\tau(v)_i = \mathcal{P}f_\tau(u)_i] \\ \geq& 1 - 5\alpha_3 - 5\alpha_3.\end{aligned}$$

□

Claim 3.21. *For every excellent τ,*

$$\Pr_{w\in[N]^{\bar{A}},i\in\bar{A}}[\mathcal{P}f_\tau(w)_i = g_\tau(w)_i] \geq 1-20\alpha_3\,.$$

Proof. Fix an excellent τ. The function g_τ is defined as the plurality of $\mathcal{P}f_\tau$. This means that for every $w\in[N]^{\bar{A}}$ and $i\in\bar{A}$, if $\mathcal{P}f_\tau(w)_i \neq g_\tau(w)_i$, then $\mathcal{P}f_\tau(w)_i$ is not the most frequent value $\mathcal{P}f_\tau(v)_i$ among all v such that $v_i = w_i$, i. e., $\Pr_{v\in[N]^{\bar{A}}}[\mathcal{P}f_\tau(w)_i = \mathcal{P}f_\tau(v)_i | w_i = v_i] \leq 1/2$.

This implies that

$$\Pr_{w,v\in[N]^{\bar{A}},i\in\bar{A}}[\mathcal{P}f_\tau(w)_i \neq \mathcal{P}f_\tau(v)_i \mid w_i = v_i] \geq \frac{1}{2}\Pr_{w\in[N]^{A},i\in\bar{A}}[\mathcal{P}f_\tau(w)_i \neq g_\tau(w)_i]\ .$$

Therefore, by Claim 3.20,

$$\Pr_{w\in[N]^{A},i\in\bar{A}}[\mathcal{P}f_\tau(w)_i \neq g_\tau(w)_i] \leq 20\alpha_3 k\,. \qquad \square$$

Proof of Lemma 3.13. Fix an excellent τ. For every $w\in[N]^{\bar{A}}$, let $D_w\subset\bar{A}$ be the set of coordinates

$$D_w = \left\{ i\in\bar{A} \,\middle|\, g_\tau(w)_i \neq \mathcal{P}f_\tau(w)_i \right\}\ .$$

Let $C\subset[N]^{\bar{A}}$ be the set of inputs in which g_τ is close to $\mathcal{P}f_\tau$,

$$C = \left\{ w\in[N]^{\bar{A}} \,\middle|\, |D_w| \leq 25\alpha_3 k \right\}\ .$$

By Claim 3.21 and averaging, $\Pr_w[w\in C]\geq 1/5$.

Let $B\subset[N]^{\bar{A}}$ be the set of inputs in which $g_\tau,\mathcal{P}f_\tau$ are far from each other,

$$B = \left\{ w\in[N]^{k'} \,\middle|\, |D_w| \geq \alpha_2 k \right\}\ .$$

Notice that $B\cap C=\phi$ but there can be inputs that are neither in B nor in C (because $\alpha_2 = 1500\alpha_0$ and $\alpha_3 = 20\alpha_0$).

By Corollary 2.9,

$$\Pr_{w,v\underset{1/2,J}{\sim}w}[w\in B, v\in C] \geq \Pr_{v\in[N]^{\bar{A}}}[v\in C]^{\frac{3}{2}}\Pr_{w\in[N]^{\bar{A}}}[w\in B]^{\frac{5}{2}} \geq \frac{1}{12}\Pr_{w\in[N]^{\bar{A}}}[w\in B]^{\frac{5}{2}}\,.$$

We will show that $\mathcal{P}f_\tau$ has large disagreement on almost every $\frac{1}{2}$-correlated pair w,v such that $w\in B, v\in C$, which lets us bound the size of B.

Typically, for $v\underset{1/2,J}{\sim}w$, we get that $|J\cap D_w|\approx|D_w|/2$. The probability of $|J\cap D_w|\geq(1/2-1/10)\,|D_w|$ is nearly 1, by the Chernoff bound, $\Pr_{v\underset{1/2,J}{\sim}w}[|J\cap D_w|<(1/2-1/10)\,|D_w|]\leq e^{\frac{1}{3\cdot 5^2}\frac{1}{2}|D_w|}$.

If $w \in B$ and $v \in C$ and $|J \cap D_w| \geq (1/2 - 1/10)\,|D_w|$, it must be that $\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\not\approx} \mathcal{P} f_\tau(v)_J$. This is because for $v \in C$, $\mathcal{P} f_\tau(v)_J \overset{25\alpha_3 k}{\approx} g_\tau(w)_J$. For $w \in B$, $|D_w| \geq \alpha_2 k$, if $|J \cap D_w| \geq (1/2 - 1/10)\,\alpha_2 k$, then $|J \cap D_w| \geq 30\alpha_3 k$, i.e., $\mathcal{P} f_\tau(w)_J \overset{30\alpha_3 k}{\not\approx} g_\tau(w)_J$. The function g_τ is a product function, so $g_\tau(w)_J = g_\tau(v)_J$. So if $\mathcal{P} f_\tau(w)_J \overset{30\alpha_3 k}{\not\approx} g_\tau(w)_J$ and $\mathcal{P} f_\tau(v)_J \overset{25\alpha_3 k}{\approx} g_\tau(w)_J$, it must be that $\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\not\approx} \mathcal{P} f_\tau(v)_J$.

From the two equations above,

$$\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}\left[w \in B, v \in C \text{ and } |J \cap D_w| \geq \left(\frac{1}{2} - \frac{1}{10}\right)|D_w|\right] \geq \frac{1}{12} \Pr_{w \in [N]^{\bar{A}}}[w \in B]^{\frac{5}{2}} - \mathrm{e}^{-\alpha_0 k}. \quad (3.12)$$

That is, $\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}\left[\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\not\approx} \mathcal{P} f_\tau(v)_J\right] \geq \frac{1}{12} \Pr_{w \in [N]^{\bar{A}}}[w \in B]^{\frac{5}{2}} - \mathrm{e}^{-\alpha k}$.

From Claim 3.19, $\Pr_{w \in [N]^{\bar{A}}, v \underset{1/2,J}{\sim} w}\left[\mathcal{P} f_\tau(w)_J \overset{\alpha_3 k}{\not\approx} \mathcal{P} f_\tau(v)_J\right] \leq 3\epsilon^{10}$. Therefore,

$$\frac{1}{12} \Pr_{w \in [N]^{\bar{A}}}[w \in B]^{\frac{5}{2}} - \mathrm{e}^{-\alpha k} \leq 3\epsilon^{10}. \quad (3.13)$$

This means that $\Pr_{w \in [N]^{\bar{A}}}[w \in B] \leq 3\epsilon^4$, and completes the proof of Lemma 3.13. □

Finally, we are left with proving the following claim, relating the standard 1/2-correlated distribution to an almost correlated version in which one extra random coordinate is identical.

Definition 3.22. For every $x \in [N]^{\bar{A}}$, we say that y is almost 1/2-correlated with x, denoted by $y \underset{1/2,J^A}{\sim} x$ i it is chosen by the following process:

- Choose a uniform $i \in \bar{A}$ and set $J = \{i\}$.
- Insert any $j \neq i$ into J with probability 1/2 independently.
- Set $y_J = x_J$, set the rest of y to be uniform.

Claim 3.23. *For any event $E(y, x, J)$ over $x, y \in [N]^{\bar{A}}$ and $J \subseteq \bar{A}$,*

$$\Pr_{y \underset{1/2,J^A}{\sim} x}[E(y,x,J)] \leq 2 \Pr_{y \in [N]^k, x \underset{1/2,J}{\sim} y}[E(y,x,J)]\,.$$

Proof. Fix $E(x, y, J)$ to be any event. For every $x \in [N]^k$, let B_x contain the tuples (x, J) such that $E(x, y, J)$ happens.

Fix $x \in [N]^k$. For each $(y, J) \in B_y$, by the definition of $\frac{1}{2}$-correlation,

$$\Pr_{z \underset{1/2,J'}{\sim} x}[(z, J') = (y, J)] = \left(\frac{1}{2}\right)^{|J|}\left(\frac{1}{2}\right)^{k-|J|}\left(\frac{1}{N}\right)^{k-|J|}.$$

For z, J' that are almost $\frac{1}{2}$-correlated to y:

$$\Pr_{z \underset{1/2,J'^A}{\sim} x}[(z, J') = (x, J)] = \frac{|J|}{k}\left(\frac{1}{2}\right)^{|J|-1}\left(\frac{1}{2}\right)^{k-|J|}\left(\frac{1}{N}\right)^{k-|J|} .$$

Therefore,

$$\Pr_{z \underset{1/2,J'^A}{\sim} y}[(z, J') = (x, J)] \leq 2 \Pr_{z \underset{1/2,J'}{\sim} y}[(z, J') = (x, J)] .$$

This implies that

$$\Pr_{y \in [N]^k, x \underset{1/2,J^A}{\sim} y}[E(y, x, J)] = \Pr_{y \in [N]^k, x \underset{1/2,J^A}{\sim} y}\left[(x, J) \in B_y\right] \leq 2 \Pr_{y \in [N]^k, x \underset{1/2,J}{\sim} y}\left[(x, J) \in B_y\right] ,$$

which finishes the proof. ◻

4 Global structure for sets

Up until now we have considered functions $f : [N]^k \to [M]^k$ whose inputs are ordered tuples $(x_1 \ldots, x_k) \in [N]^k$. We now move to consider functions $f : \binom{[N]}{k} \to [M]^k$ whose inputs are unordered sets $\{x_1, \ldots, x_k\} \in \binom{[N]}{k}$. In this setting assume that $N \gg k$ (for tuples no such assumption is made).

To each subset $S = \{s_1, \ldots, s_k\}$ the function f assigns $f(S) \in [M]^k$. We view $f(S)$ as a "local function" on S, assigning a value in $[M]$ to every $a \in S$. We denote by $f(S)_a$ the value that $f(S)$ assigns to a. For a subset $W \subset S$, we denote by $f(S)_W$ the outputs of f corresponding to the elements in W.

There are straightforward analogs to Theorem 1.1 and Theorem 3.9 which we present and prove in this section. Interestingly, in the case of sets deducing global structure from restricted global structure is quite easier than it is for tuples.

First, let us present the Z-test for sets from [13] when $t = k/10$. Let $\mathrm{agr}^{Z_{set}}_{k/10}(f)$ be the success probability of this test. This is the same test as Test 3 from the introduction with $t = k/10$.

For convenience, we write again Theorem 1.5 from the introduction.

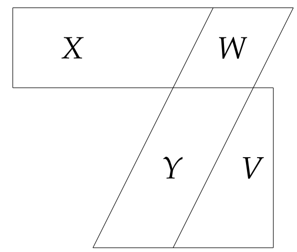
Theorem 4.1 (Global Structure for Sets, restated)**.** *There exist a small constant $c > 0$, such that for every constant $\lambda > 0$, large enough $k \in \mathbb{N}$ and $N > \mathrm{e}^{c\lambda k}$, $M \in \mathbb{N}$, if the function $f : \binom{[N]}{k} \to [M]^k$ passes Test 3 with probability $agr^{Z_{set}}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\lambda k}$, then there exists a function $g : [N] \to [M]$ such that*

$$\Pr_S\left[f(S) \overset{\lambda k}{\approx} g(S)\right] \geq \epsilon - 4\epsilon^2 .$$

To prove the theorem, we first "translate" the restricted global structure theorem for tuples to a theorem on sets, and then use it to prove the global structure.

Test 4: Z-test for functions over sets, with $t = k/10$ (3-query test)

1. Choose a random set $W \subset [N]$ of size $k/10$.
2. Choose random $V, W, X, Y \subset [N]$, such that $|W| = |V| = k/10$, $|X| = |Y| = 9k/10$ and $X \cap W = Y \cap W = Y \cap V = \emptyset$.
3. Accept if $f(X \cup W)_W = f(Y \cup W)_W$ and $f(Y \cup W)_Y = f(Y \cup V)_Y$.



Denote by $agr^{Z_{set}}_{k/10}(f)$ the success probability of f on this test.

4.1 Restricted global structure for sets

We define analogous definitions for good restrictions and DP restrictions for functions on sets. To make the reduction proof simpler, we use a constant $\eta \in \left[1 - k^2/N, 1\right]$ (i. e., almost 1). Fix a function $f : \binom{[N]}{k} \to [M]^k$ such that $\mathrm{agr}^{Z_{set}}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\lambda k}$.

Definition 4.2 (Good pair)**.** A pair $X, W \subset [N]$, $|X| = 9k/10$, $|W| = k/10$, $X \cap W = \emptyset$ is good if

$$\Pr_Y \left[f(X \cup W)_W = f(Y \cup W)_W \mid Y \cap W = \emptyset \right] > \frac{\epsilon}{2}\eta .$$

Definition 4.3 (α-DP pair)**.** A pair $X, W \subset [N]$, $|X| = 9k/10$, $|W| = k/10$, $X \cap W = \emptyset$ is an α-DP pair if it is good, and if there exists a function $g_{X,W} : [N] \to [M]$ such that

$$\Pr_Y \left[f(Y \cup W)_Y \overset{3\alpha k}{\not\approx} g_{X,W}(Y) \,\middle|\, Y \cap W = \emptyset, f(X \cup W)_W = f(Y \cup W)_W \right] \le 2\epsilon^2 .$$

Notice that in the case of a function on sets, there is a single function $g_{X,W} : [N] \to [M]$, instead of $9k/10$ different functions in the case of tuples.

Lemma 4.4 (Restricted global structure for sets)**.** *There exists a small constants $c > 0$, such that for every constant $\alpha > 0$, large enough $k \in \mathbb{N}$ and $N > \mathrm{e}^{c\lambda k}$, $M \in \mathbb{N}$, the following holds.*

For every function $f : \binom{[N]}{k} \to [M]^k$, if $agr^{Z_{set}}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\alpha k}$, then at least $(1 - \epsilon^2 - k^2/N)$ of the good pairs W, X are α-DP pairs.

This lemma is an analog to Theorem 3.9, and we prove it by a reduction from it. For every $f : \binom{[N]}{k} \to [M]^k$ we define a function $f' : [N]^k \to [M]^k \cup \perp$ that equals $\perp$ if the input has two identical coordinates, and identifies with f everywhere else. For $N \gg k$, almost all inputs don't have two identical coordinates, and f', f are equal on these inputs.

Using Theorem 3.9, we derive a restricted global structure on f'. Since f equals f' almost always, we find an equivalence between an α-DP pair X, W to an α-DP restriction τ. For every α-DP restriction τ we have the direct product function $g_\tau = (g_i)_{i \in \bar{A}}, g_i : [N] \to [M]$. We build a restricted global function $g_{X,W} : [N] \to [M]$ by taking the most frequent value among the functions $(g_i)_{i \in \bar{A}}$. Note that even though f' is permutation invariant, the functions $(g_i)_{i \in \bar{A}}$ are not necessarily identical.

Since the proof is technical, and its main points are described in the paragraph above, we defer it to Appendix A.

4.2 Global structure for sets

The proof is very similar to the proof of lemma 3.16 in [13].

Proof of Theorem 1.5. Fix a function $f : \binom{[N]}{k} \to [M]^k$ that passes Test 4 with probability ϵ, denote $\alpha = \lambda/5$. Let $W, X \subset [N]$ be the subsets chosen on Test 4.

If $\Pr_Y[f(X \cup W)_W = f(Y \cup W)_W \mid Y \cap W = \emptyset] < \frac{\epsilon}{2}\eta$, then the test rejects with probability at least $1 - \frac{\epsilon}{2}\eta$. The function f passes the test with probability ϵ, so the test must succeed with probability at least ϵ on W, X such that $\Pr_Y[f(X \cup W)_W = f(Y \cup W)_W \mid Y \cap W = \emptyset] > \frac{\epsilon}{2}\eta$, i. e., on good pairs.

Using Lemma 4.4, at least $(1 - 2\epsilon^2 - k^2/N)$ of the good pairs are α-DP pairs. Fix an α-DP pair W, X, and let $g = g_{X,W} : [N] \to [M]$ be the direct product function associated with W, X. Let $\mathcal{V}$ be all the sets Y that are consistent with W, X,

$$\mathcal{V} = \left\{ Y \in \binom{[N]}{9k/10} \,\middle|\, Y \cap W = \emptyset, f(X \cup W)_W = f(Y \cup W)_W \right\}.$$

We use the third query to show that this g is in fact a global direct product function which is close to f, i.e that $f(S) \approx g(S)$ for about an ϵ-fraction of the sets $S \in \binom{[N]}{k}$.

Let C be the set of inputs for which f, g are close,

$$C = \left\{ S \in \binom{[N]}{k} \,\middle|\, f(S) \overset{\lambda k}{\approx} g(S) \right\}.$$

Suppose that instead of running Test 4 as is, we choose Y, V by the following process:

1. Choose a uniform $S \in \binom{[N]}{k}$.
2. Choose Y to be a uniform $9k/10$ subset of S.
3. Set $V = S \setminus Y$ and return (Y, V).

If the process outputs Y such that $Y \cap W \neq \emptyset$, we assume that the test rejects. The probability of this event is less than k^2/N, and if it does not happen, the process generates the test distribution. Therefore, f passes the test where Y, V are chosen using the above process with probability at least $\epsilon - k^2/N$. We prove the claim by proving that conditioning on $S \notin C$, the success probability of the test is much lower than ϵ. We deduce that the probability of $S \in C$ must be close to ϵ.

The test passes if $Y \in \mathcal{V}$ and $f(Y \cup W)_Y = f(Y \cup V)_Y$. We prove that for $S \notin C$, with high probability both $f(Y \cup V)_Y \overset{3\alpha k}{\not\approx} g(Y)$ and $f(Y \cup W)_Y \overset{3\alpha k}{\approx} g(Y)$, and the test fails.

1. If $Y \notin \mathcal{V}$, the test fails. For every $Y \in \mathcal{V}$, Lemma 4.4 states that $f(Y \cup W)_Y \overset{3\alpha k}{\approx} g(Y)$ with probability $1 - 2\epsilon^2$. Conditioning on $S \notin C$ does not increase the probability by much (where we assume that C is small, else there is nothing to prove),

$$\Pr_Y\left[f(Y \cup W)_Y \overset{3\alpha k}{\not\approx} g(Y) \,\middle|\, Y \in \mathcal{V}, S \notin C\right] \leq 2\epsilon^2 \frac{1}{\Pr[S \notin C]} \leq 3\epsilon^2 . \tag{4.1}$$

2. For every set S let $D_S = \{i \in [k] \mid f(S)_i \neq f(S)_i\}$. For $S \notin C$, $|D_S| > 5\alpha k$. The set Y is a random subset of S of size $9k/10$. Therefore, for $S \notin C$, by the tail bound (Fact 2.4), $\Pr_{Y \subset S}[|Y \cap D_S| \leq 3\alpha k \mid] \leq e^{-\frac{\alpha k}{4}}$. By definition, if $|Y \cap D_S| \geq 3\alpha k$, then $f(S)_Y \overset{3\alpha k}{\not\approx} g(Y)$. That is, for $S \notin C$, $\Pr_{Y \subset S}\left[f(S)_Y \overset{3\alpha k}{\approx} g(Y)\right] \leq e^{-\frac{\alpha k}{4}}$.

From the two items above,

$$\begin{aligned}
\Pr[\text{Test passes}|S \notin C] &= \Pr\left[Y \in \mathcal{V} \text{ and } f(W \cup Y)_Y = f(V \cup Y)_Y \mid S \notin C\right] \\
&\leq \Pr_Y\left[f(Y \cup W)_Y \overset{3\alpha k}{\not\approx} g(Y) \text{ and } Y \in \mathcal{V} \,\middle|\, S \notin C\right] \\
&+ \Pr_{Y \subset S}\left[f(S)_Y \overset{3\alpha k}{\approx} g(Y) \,\middle|\, S \in C\right] \\
&\leq 3\epsilon^2 + e^{-\frac{\alpha k}{4}} .
\end{aligned}$$

We assume that f passes the test with probability $\epsilon - k^2/N$,

$$\Pr[\text{Test passes}] = \Pr[\text{Test passes and } S \in C] + \Pr[\text{Test passes and } S \notin C] .$$

Therefore,

$$\Pr[S \in C] \geq \epsilon - \frac{k^2}{N} - 3\epsilon^2 - e^{-\frac{1}{4}\alpha k} \geq \epsilon - 4\epsilon^2 ,$$

which finishes the proof. □

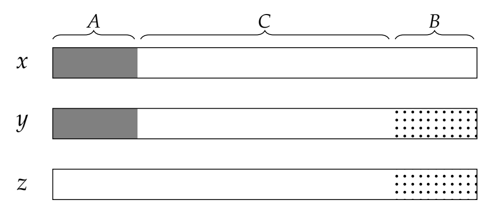
In the introduction, we stressed that in order to extend the restricted global structure into a global structure, the restricted global structure theorem has to be "strong," i. e., the probability of $f(Y \cup W)_Y \overset{3\alpha k}{\not\approx} g_{X,W}(Y)$ should be strictly smaller than ϵ, it is $2\epsilon^2$ in our case. If the local structure was not strong and the bound in (4.1) would have been larger than ϵ, then all the success probability of the test could come from sets such that $f(S) \overset{5\alpha k}{\not\approx} g(S)$. In this case, we could not have deduced that C is large and g is close to a direct product function. This is the situation in the local structure theorem of [8].

5 Global structure for tuples

In this section, we prove our main theorem – global structure for tuples. The proof uses the restricted global structure, Theorem 3.9. For convenience, we write again the test and theorem from the introduction.

Test 1: Z-test with parameter $t = k/10$ (3-query test)

1. Choose A, B, C to be a random partition of $[k]$, such that $|A| = |B| = k/10$.
2. Choose uniformly at random $x, y, z \in [N]^k$ such that $x_A = y_A$ and $y_B = z_B$.
3. Accept if $f(x)_A = f(y)_A$ and $f(z)_B = f(y)_B$.



Denote by $agr^Z_{k/10}(f)$ the success probability of f on this test.

Let $\mathcal{D}^z$ be the distribution over A, B, C, x, y, z as described above in the test.

Theorem 5.1 (Main theorem – Global Structure for tuples, restated)**.** *For every $N, M > 1$, there exists a small constant $c > 0$ such that for every constant $\lambda > 0$ and large enough k, if $f : [N]^k \to [M]^k$ is a function that passes Test 1 with probability $\epsilon = agr^Z_{k/10}(f) \geq \mathrm{e}^{-c\lambda^2 k}$, then there exist functions $(g_1, \dots, g_k)$, $g_i : [N] \to [M]$ such that*

$$\Pr_{x \in [N]^k}\left[f(x) \overset{\lambda k}{\approx} (g_1(x_1), \dots, g_k(x_k))\right] \geq \frac{\epsilon}{10}.$$

Fix a function $f : [N]^k \to [M]^k$, such that $\mathrm{agr}^Z_{k/10}(f) = \epsilon \geq \mathrm{e}^{-c\lambda^2 k}$.

Similar to the proof of the restricted global structure, in the proof we use several values for the slack parameter λ. These are constant multiples of each other, and are denoted by λ_0, λ_1 etc.

Our proof of Theorem 1.1 relies on Theorem 3.9. The theorem states that many restrictions τ are DP restrictions (see Definition 3.6). Fix $\lambda_0 = \lambda/10000$. We apply theorem Theorem 3.9 with the slack parameter $\alpha = \lambda_0$. For every λ_0-DP restriction τ, we denote by g_τ the direct product function corresponding to τ.

For the proof, we need a few definitions.

Definition 5.2 (Successful set)**.** For every $x \in [N]^k$, a set $A \subset [k], |A| = k/10$, is *successful* with respect to x if

1. $\Pr_{y,B,z \sim \mathcal{D}^z_{|A,x}}$ [Test 1 succeeds] $\geq \epsilon/3$.
2. $\tau = (A, x_A, f(x)_A)$ is a λ_0-DP-restriction.

Definition 5.3 (Consistent functions)**.** Let $S_1, S_2 \subset [k]$ and let $g : [N]^{S_1} \to [M]^{S_1}$ and $g' : [N]^{S_2} \to [M]^{S_2}$ be two direct product functions. We say that g, g' are β-consistent if for a uniform $i \in S_1 \cap S_2$ and $b \in [N]$,

$$\Pr_{i,b}\left[g_i(b) \neq g'_i(b)\right] \leq \beta\,.$$

We prove the main theorem in Section 5.1, and prove the lemmas used in the proof in the next sections.

5.1 Proof of Theorem 1.1

Let $f : [N]^k \to [M]^k$ be a function that passes Test 1 with probability ϵ.

By Theorem 3.9, a restriction $\tau = (A, x_A, f(x)_A)$ is a λ_0-DP-restriction with probability at least $\epsilon/2 \cdot (1-\epsilon^2)$. We denote by $g_\tau : [N]^{\bar{A}} \to [M]^{\bar{A}}$ the DP function associated with τ.

We start by finding a single string x which is "globally good". Fix $\lambda_1 = 60\lambda_0$.

Lemma 5.4. *There exists $x \in [N]^k$, such that*

1. $\Pr_A$ *[A is successful with respect to x]* $\geq \epsilon/4$.
2. $\Pr_{A_1,A_2 \in \binom{[k]}{k/10}}$ *[A_1, A_2 are successful w.r.t. x and g_{τ_1}, g_{τ_2} are λ_1-consistent]* $\geq \epsilon^5$*, where the tuples are $\tau_1 = (A_1, x_{A_1}, f(x)_{A_1})$ and $\tau_2 = (A_2, x_{A_2}, f(x)_{A_2})$.*

The proof appears on Section 5.2.

We fix the string $x \in [N]^k$ promised from Lemma 5.4. Let $\mathcal{A} \subset \binom{[k]}{k/10}$ be the set of sets that are successful with respect to x. For every $A \in \mathcal{A}$, let $g_A : [N]^{\bar{A}} \to [M]^{\bar{A}}$ be the direct product function g_τ, for $\tau = (A, x_A, f(x)_A)$.

Theorem 5.5. *For all integers $N, M \in \mathbb{N}$ and large enough $k \in \mathbb{N}$, and all small constants $\beta, \nu > 0$ such that $\nu > \mathrm{e}^{-\frac{1}{3}\beta^2 k}$, the following holds. Let $\mathcal{A} \subset \binom{[k]}{k/10}$ be a set of sets, and let $\mathcal{F} = \{g_A : [N]^{\bar{A}} \to [M]^{\bar{A}}\}_{A \in \mathcal{A}}$ be a family of direct product functions. If*

$$\Pr_{A_1,A_2 \in \binom{[k]}{k/10}}\left[A_1, A_2 \in \mathcal{A} \text{ and } g_{A_1}, g_{A_2} \text{ are } \beta\text{-consistent}\right] \geq \nu\,,$$

then there exists a global function $g : [N]^k \to [M]^k$ such that

$$\Pr_{A \in \binom{[k]}{k/10}}\left[A \in \mathcal{A} \text{ and } g_A, g \text{ are } 50\beta\text{-consistent}\right] \geq \frac{1}{4}\nu\,.$$

We prove the theorem in Section 5.3 (in fact we prove a slightly stronger statement).

The family $\{g_A\}_{A \in \mathcal{A}}$ satisfies the conditions of the theorem, with parameters $\nu = \epsilon^5$ and $\beta = \lambda_1$. Let $g : [N]^k \to [M]^k$ be the direct product function promised from the theorem. Fix $\lambda_3 = 160\lambda_1$. We finally claim that g is the required global DP function,

Lemma 5.6. $\Pr_{z \in [N]^k}\left[f(z) \overset{\lambda_3 k}{\approx} g(z)\right] \geq \frac{\epsilon}{10}$.

The proof appears in Section 5.4. This finishes the proof since $\lambda_3 < \lambda$.

5.2 Consistency between restricted global functions

In this section we find a string $x \in [N]^k$ which is "globally good", proving Lemma 5.4.

Claim 5.7. *There exists $x \in [N]^k$ such that*

1. $\Pr_{A,y,B,z\sim\mathcal{D}^z_{|x}}$ [*Test 1 passes*] $\geq \epsilon/3$ *and,*
2. $\Pr_A$ [*A is successful w.r.t. x*] $\geq \frac{\epsilon}{3}$.

Proof. Let $G = (L \cup R, E)$ be the full bipartite graph, with vertex sets $L = \binom{[k]}{k/10}$ and $R = [N]^k$. Let $\omega : E \longrightarrow [0,1]$ be a function matching each edge (A, x) the success probability of Test 1 given that A, x are chosen, i. e., $\omega(A, x) = \Pr_{y,B,z\sim\mathcal{D}^z_{|A,x}}$[Test 1 passes]. Then $\mathbb{E}_{(A,x)\in E}[\omega(A, x)] \geq \epsilon$.

Suppose that for each edge e with $\omega(e) \leq \epsilon/2$, we change $\omega(e)$ to 0. The expected value of ω is reduced by at most $\epsilon/2$, i. e., $\mathbb{E}_{(A,x)\in E}[\omega(A, x)] \geq \epsilon/2$. All the edges (A, x) that remain with positive value ω are of edges (A, x) such that $\tau = (A, x, f(x)_A)$ is good[2].

We further change $\omega(e)$ to 0 for all edges $e = (A, x)$ such that $\tau = (A, x_A, f(x)_A)$ is not a λ_0-DP restriction. From Theorem 3.9, a good $\tau \sim \mathcal{D}$ is a DP-restriction with probability at least $1 - \epsilon^2$. The distribution $\tau \sim \mathcal{D}$ corresponds to a uniform choice of $(A, x) \in E$. Therefore, we have changed $\omega(e)$ to 0 on at most ϵ^2 fraction of the edges. The maximal value of ω is 1, so this step reduces the expectation of ω over E by at most ϵ^2, and $\mathbb{E}_{(A,x)\in E}[\omega(A, x)] \geq \epsilon/2 - \epsilon^2 \geq \epsilon/3$.

Let x be a vertex which maximizes $\mathbb{E}_A[\omega(A, x)]$, then

$$\Pr_{A,y,B,z\sim\mathcal{D}^z_{|x}}\left[\text{Test 1 passes}\right] \geq \mathop{\mathbb{E}}_A[\omega(A, x)] \geq \frac{\epsilon}{3}.$$

All edges (A, x) such that $\omega(A, x) > 0$ are such that A is successful with respect to x, so

$$\Pr_A[A \text{ is successful w.r.t. } x] = \Pr_A[\omega(A, x) > 0] \geq \frac{\epsilon}{3},$$

where the last inequality holds because the maximal value of ω is 1. $\square$

In the rest of this subsection, we fix an x satisfying the properties of Claim 5.7. For every set A that is successful with respect to x, denote by $g_A : [N]^{\bar{A}} \longrightarrow [M]^{\bar{A}}$ the function g_τ for $\tau = (A, x_A, f(x)_A)$. Denote by $\mathcal{Z}_A$ the set

$$\mathcal{Z}_A = \left\{ z \in [N]^k \;\middle|\; f(z)_{\bar{A}} \overset{\frac{\lambda_1 k}{3}}{\approx} g_A(z_{\bar{A}}) \right\}.$$

Note that the set $\mathcal{Z}_A$ might be all of $[N]^k$.

We prove in the next claim that if A is successful with respect to x, then g_A is consistent with the original function f. This consistency is much stronger than what is guaranteed by Theorem 3.9. By Theorem 3.9, g_A, f are consistent on a set of inputs contained in the set $\left\{ w \in [N]^k \;\middle|\; w_A = x_A \right\}$. In the claim below, we prove that g_A, f are consistent on $\Omega(\epsilon)$ fraction of $[N]^k$, which is a much larger set.

[2]There may also be good tuples with value 0, if Test 2 passes with probability larger than $\epsilon/2$ but Test 1 doesn't.

Claim 5.8. *For every A that is successful with respect to x,*

$$\Pr_{z\in[N]^k}[z\in\mathcal{Z}_A]\geq\frac{\epsilon}{4}\,.$$

Proof. Fix an A that is successful with respect to x, and assume for a contradiction that $\Pr_{z\in[N]^k}[z\in\mathcal{Z}_A]<\epsilon/4$. For every z, let $D_z=\left\{i\in\bar{A}\,\middle|\, f(z)_i\neq g_{A,i}(z_i)\right\}$, then for $z\notin\mathcal{Z}_A$, $|D_z|\geq k\lambda_1/3$.

Let $y,z,B\sim\mathcal{D}^{\mathrm{z}}|A,x$. Suppose y is such that $f(y)_{\bar{A}}\overset{\lambda_0 k}{\approx}g_A(y_{\bar{A}})$. The test passes if $f(y)_B=f(z)_B$, which implies (since $B\subset\bar{A}$) that also $f(z)_B\overset{\lambda_0 k}{\approx}g_A(y_{\bar{A}})_B$. The function g_A is a product function and $y_B=z_B$, so $g_A(y_{\bar{A}})_B=g_A(z_{\bar{A}})_B$ and we get from the previous inequality that

$$f(z)_B\overset{\lambda_0 k}{\approx}g_A(y_{\bar{A}})_B=g_A(z_{\bar{A}})_B.$$

From the definition of D_z, this only happens if $|D_z\cap B|\leq\lambda_0 k$. To summarize, if the test passes and y is such that $f(y)_{\bar{A}}\overset{\lambda_0 k}{\approx}g_A(y_{\bar{A}})$, it must be that $|D_z\cap B|\leq\lambda_0 k$.

From the above paragraph,

$$\begin{aligned}\Pr_{B,y,z\sim\mathcal{D}^{\mathrm{z}}_{|A,x}}[\text{Test passes}|z\notin\mathcal{Z}_A]&=\Pr_{B,y,z}\left[\text{Test passes and } f(y)_{\bar{A}}\overset{\lambda_0 k}{\not\approx}g_A(y_{\bar{A}})\,\middle|\, z\notin\mathcal{Z}_A\right]\\&\quad+\Pr_{B,y,z}\left[\text{Test passes and } f(y)_{\bar{A}}\overset{\lambda_0 k}{\approx}g_A(y_{\bar{A}})\,\middle|\, z\notin\mathcal{Z}_A\right]\\&\leq\Pr_y\left[f(y)_{\bar{A}}\overset{\lambda_0 k}{\not\approx}g_A(y_{\bar{A}})\,\middle|\, f(x)_A=f(y)_A,z\notin\mathcal{Z}_A\right]\\&\quad+\Pr_{B,z}[|B\cap D_z|\leq\lambda_0 k\mid z\notin\mathcal{Z}_A]\,.\end{aligned}$$

We bound the two expressions. For the first, from Theorem 3.9,

$$\Pr_y\left[f(y)_{\bar{A}}\overset{\lambda_0 k}{\not\approx}g_A(y_{\bar{A}})\,\middle|\, f(x)_A=f(y)_A\right]\leq\epsilon^2.$$

Conditioning on $z\notin\mathcal{Z}_A$, which occurs with probability at least $1-\frac{\epsilon}{4}$, increases the probability by a factor of at most $\frac{1}{1-\frac{\epsilon}{4}}<2$.

For the second expression, the set B is a random subset of $\bar{A}$ of size $k/10$, and $|D_z|\geq k\lambda_1/3=20\lambda_0 k$. Using the Hoeffding bound for random subset (Fact 2.4),

$$\Pr_{B,z}[|B\cap D_z|\leq\lambda_0 k\mid z\notin\mathcal{Z}_A]\leq\mathrm{e}^{-\frac{\lambda_1 k}{20}}<\epsilon^2\,.$$

We conclude that

$$\Pr_{B,y,z\sim\mathcal{D}^{\mathrm{z}}_{|A,x}}[\text{Test passes}|z\notin\mathcal{Z}_A]\leq 3\epsilon^2\,.$$

This implies that

$$\begin{aligned}\Pr_{B,y,z\sim\mathcal{D}^z_{|A,x}}[\text{Test passes}] \leq & \Pr_{B,y,z\sim\mathcal{D}^z_{|A,x}}[\text{Test passes}|z\notin\mathcal{Z}_A]+\Pr_z[z\in\mathcal{Z}_A] \\ \leq & 3\epsilon^2+\frac{\epsilon}{4}<\frac{\epsilon}{3}\,.\end{aligned}$$

This contradicts A being successful with respect to x. ◻

In the introduction, we explained the difference between our restricted global structure and the result of [8]. In our result, Theorem 3.9, $f(y)_{\bar{A}} \approx g_A(y)$ for $1-\epsilon^2$ of $y\in\mathcal{V}_\tau$ (for $\tau=(A,x_A,f(x)_A)$, whereas in their result this probability was not as overwhelmingly close to 1. We require this for proving the above claim, as well as for proving the global structure.

Claim 5.9.

$$\Pr_{A_1,A_2\in\binom{[k]}{k/10}}\left[|\mathcal{Z}_{A_1}\cap\mathcal{Z}_{A_2}|\geq\frac{\epsilon^2}{32}N^k\;\middle|\;A_1,A_2\textit{ are successful w.r.t. } x\right]\geq\frac{\epsilon^2}{32}\,.$$

Proof. Let A_1,A_2 be two uniform sets that are successful with respect to x, then

$$\begin{aligned}\mathop{\mathbb{E}}_{A_1,A_2}[|\mathcal{Z}_{A_1}\cap\mathcal{Z}_{A_2}|] &= \sum_{z\in[N]^k}\mathop{\mathbb{E}}_{A_1,A_2}[\mathbb{I}(z\in\mathcal{Z}_{A_1}\cap\mathcal{Z}_{A_2})] \\ &= \sum_{z\in[N]^k}\mathop{\mathbb{E}}_{A_1,A_2}[\mathbb{I}(z\in\mathcal{Z}_{A_1})\mathbb{I}(z\in\mathcal{Z}_{A_2})] \\ &= \sum_{z\in[N]^k}\mathop{\mathbb{E}}_{A}[\mathbb{I}(z\in\mathcal{Z}_{A})]^2\,,\end{aligned}$$

where $\mathbb{I}$ is an indicator. The last equality holds since A_1,A_2 are independent uniform sets that are successful with respect to x.

By Cauchy Schwarz,

$$\begin{aligned}\left(\sum_{z\in[N]^k}N^{-\frac{k}{2}}\mathop{\mathbb{E}}_A[\mathbb{I}(z\in\mathcal{Z}_A)]\right)^2 \leq & \left(\sum_{z\in[N]^k}N^{-k}\right)\left(\sum_{z\in[N]^k}\mathop{\mathbb{E}}_A[\mathbb{I}(z\in\mathcal{Z}_A)]^2\right) \\ = & 1\cdot\sum_{z\in[N]^k}\mathop{\mathbb{E}}_A[\mathbb{I}(z\in\mathcal{Z}_A)]^2\,.\end{aligned}$$

From Claim 5.8, for *every* A which is successful with respect to x, $\Pr_z[z\in\mathcal{Z}_A]\geq\epsilon/4$. This means that for every such A, $\sum_z\mathbb{I}(z\in\mathcal{Z}_A)\geq N^k\epsilon/4$. This also holds for a uniform A that is successful with respect to x. Combining it together with the above equations,

$$\mathop{\mathbb{E}}_{A_1,A_2}[|\mathcal{Z}_{A_1}\cap\mathcal{Z}_{A_2}|]=\sum_{z\in[N]^k}\mathop{\mathbb{E}}_A[\mathbb{I}(z\in\mathcal{Z}_A)]^2\geq\left(\sum_z N^{-\frac{k}{2}}\mathop{\mathbb{E}}_A[\mathbb{I}(z\in\mathcal{Z}_A)]\right)^2\geq\left(\frac{\epsilon}{4}N^{\frac{k}{2}}\right)^2\,.$$

The maximal value of $|\mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}|$ is N^k, therefore by averaging

$$\Pr_{A_1,A_2}\left[|\mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}| \geq \frac{\epsilon^2}{32}N^k\right] \geq \frac{\epsilon^2}{32}\,.$$

□

Claim 5.10. *For every $A_1, A_2 \in \binom{[k]}{k/10}$ such that $|\mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}| \geq N^k\epsilon^2/32$, the functions g_{A_1}, g_{A_2} are λ_1-consistent.*

Proof. Fix A_1, A_2 such that $|\mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}| \geq N^k\epsilon^2/32$. Let $S = [k] \setminus \{A_1 \cup A_2\}$. S is the set of coordinates both g_{A_1}, g_{A_2} are defined on, and $|S| \geq 0.8k$.

Assume for a contradiction that g_{A_1}, g_{A_2} are not λ_1-consistent, i. e.,

$$\Pr_{i\in S, b\in[N]}[g_{A_1,i}(b) \neq g_{A_2,i}(b)] > \lambda_1\,.$$

By the Chernoff tail bound (Fact 2.3),

$$\Pr_{z\in[N]^k}\left[g_{A_1}(z_{\bar{A}_1})_S \overset{\frac{2}{3}\lambda_1 k}{\approx} g_{A_2}(z_{\bar{A}_2})_S\right] \leq \mathrm{e}^{-\frac{1}{10}\lambda_1 k}. \tag{5.1}$$

We can use the Chernoff bound on the different coordinates $i \in S$ because the functions g_{A_1}, g_{A_2} are direct product functions, so their output on different coordinates is independent.

Any input $z \in \mathcal{Z}_1 \cap \mathcal{Z}_2$ satisfies both $f(z)_{\bar{A}_1} \overset{\frac{1}{3}\lambda_1 k}{\approx} g_{A_1}(z_{\bar{A}_1})$ and $f(z)_{\bar{A}_2} \overset{\frac{1}{3}\lambda_1}{\approx} g_{A_2}(z_{\bar{A}_2})$ which implies that $g_{A_1}(z_S) \overset{\frac{2}{3}\lambda_1 k}{\approx} g_{A_2}(z_S)$. That is,

$$\Pr_{z\in[N]^k}\left[g_{A_1}(z_{\bar{A}_1})_S \overset{\frac{2}{3}\lambda_1 k}{\approx} g_{A_2}(z_{\bar{A}_2})_S\right] \geq \Pr_{z\in[N]^k}[z \in \mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}] \geq \frac{\epsilon^2}{32},$$

which contradicts (5.1). □

Proof of Lemma 5.4. Let $x \in [N]^k$ be the input promised from Claim 5.7.

A set A is successful with respect to x with probability at least $\frac{\epsilon}{3}$. From Claim 5.9,

$$\Pr_{A_1,A_2\in\binom{[k]}{k/10}}\left[|\mathcal{Z}_{A_1} \cap \mathcal{Z}_{A_2}| \geq \frac{\epsilon^2}{32}N^k \,\middle|\, A_1, A_2 \text{ are successful w.r.t. } x\right] \geq \frac{\epsilon^2}{32}\,.$$

By Claim 5.10, such sets A_1, A_2 are λ_1 consistent, i. e.,

$$\Pr_{A_1,A_2\in\binom{[k]}{k/10}}\left[g^{A_1}, g^{A_2} \text{ are } \lambda_1 \text{ consistent} \,\middle|\, A_1, A_2 \text{ are successful w.r.t. } x\right] \geq \frac{\epsilon^2}{32}\,.$$

Therefore, the probability of A_1, A_2 to be successful with respect to x and λ_1 consistent is at least $\left(\frac{\epsilon}{4}\right)^2 \frac{\epsilon^2}{32} > \epsilon^5$. □

5.3 Agreement theorem in the dense case

In this section we prove Theorem 5.5, which is an agreement theorem for functions. In fact, we prove a more general version of the theorem, Theorem 5.13 below.

Let Σ be an alphabet with a distance function $\mathrm{dist} : \Sigma \times \Sigma \to [0,1]$ that satisfies the triangle inequality, i. e., $\mathrm{dist}(x,y) + \mathrm{dist}(y,z) \geq \mathrm{dist}(x,z)$ for all $x,y,z \in \Sigma$. We look at a family of functions $\mathcal{F} = \{f_S : S \to \Sigma\}_{S \in \binom{[k]}{9k/10}}$.

Definition 5.11. The difference between $f_{S_1}, f_{S_2} \in \mathcal{F}$, denoted by $\Delta(f_{S_1}, f_{S_2})$, is defined by the equation

$$\Delta(f_{S_1}, f_{S_2}) = \mathop{\mathbb{E}}_{i \in S_1 \cap S_2}[\mathrm{dist}(f_{S_1}(i), f_{S_1}(i))]\,.$$

The difference between $f_S \in \mathcal{F}$ and a function $g : [k] \to \Sigma$ is defined by the equation

$$\Delta(f_S, g) = \mathop{\mathbb{E}}_{i \in S}[\mathrm{dist}(f_S(i), g(i))]\,.$$

Definition 5.12. The agreement of the collection of local functions $\mathcal{F}$ regarding the uniform distribution with parameter β, denoted by $\mathrm{agree}_\beta(\mathcal{F})$ is defined by the equation

$$\mathrm{agree}_\beta(\mathcal{F}) = \Pr_{f_{S_1}, f_{S_2} \in \mathcal{F}}[\Delta(f_{S_1}, f_{S_2}) < \beta]\,.$$

Theorem 5.13. *For every small constant $\beta \in (0,1)$, large enough $k \in \mathbb{N}$, every $\nu > \mathrm{e}^{-\frac{1}{3}\beta^2 k}$, and every alphabet Σ with a distance measure $dist : \Sigma \times \Sigma \to [0,1]$ the following holds.*

If a collection of local functions $\mathcal{F} = \{f_S : S \to \Sigma\}_{S \in \binom{[k]}{9k/10}}$ has $\mathrm{agree}_\beta(\mathcal{F}) > \nu$, then there exists a global function $g : [k] \to \Sigma$ such that

$$\Pr_{S \in \binom{[k]}{9k/10}}[\Delta(f_S, g) \leq 50\beta] \geq \frac{1}{4}\nu\,.$$

Claim 5.14. *Theorem 5.13 implies Theorem 5.5.*

Proof. Let $N, M, k \in \mathbb{N}$, and let $\mathcal{A} \subset \binom{k]}{k/10}$ be a set of sets. Let $\{g_A\}_{A \in \mathcal{A}}, \forall A, g_A : [N]^{\bar{A}} \to [M]^{\bar{A}}$ be a family of direct product functions satisfying the conditions of Theorem 5.5.

Set $\Sigma = [M]^N \cup \perp$. For every $A \in \mathcal{A}$ and every direct product function $g_A : [N]^{\bar{A}} \to [M]^{\bar{A}}$, $g_A = (g_i)_{i \in \bar{A}}$, we define $f_{\bar{A}} : \bar{A} \to [M]^N$ by

$$\forall i \in \bar{A} \quad f_{\bar{A}}(i) = \text{ the truth table of } g_i\,.$$

Since $g_i : [N] \to [M]$, its truth table is in $[M]^N$. For every $A \notin \mathcal{A}$ we set $f_{\bar{A}}$ to equal $\perp$ on all inputs.

We define the distance measure inside Σ as follows,

$$\forall \sigma, \sigma' \in \Sigma \quad \mathrm{dist}(\sigma, \sigma') = \begin{cases} \Pr_{i \in [N]}[\sigma_i \neq \sigma'_i] & \text{if } \sigma, \sigma' \neq \perp \\ 1 & \text{otherwise.} \end{cases}$$

The distance is the normalized Hamming distance for strings that are not $\perp$.

We show that the collection of functions $\{f_S\}_{S\in\binom{[k]}{9k/10}}$ satisfies the conditions of Theorem 5.13. If $A_1, A_2 \in \mathcal{A}$ and g_{A_1}, g_{A_2} are β-consistent, then by definition $\Delta(f_{\bar{A}_1}, f_{\bar{A}_2}) < \beta$, thus by the assumptions of Theorem 5.5, $\mathrm{agree}_\beta(\mathcal{F}) > \nu$.

Let $g : [k] \to \Sigma$ be the function promised from Theorem 5.13. Then $\Delta(f_S, g) \leq 50\beta$ for at least a $\nu/4$ fraction of the sets $S \in \binom{[k]}{9k/10}$, let this set of functions be $\mathcal{F}'$. Every function $f_S \in \mathcal{F}'$ must correspond to some $g_{\bar{S}}$ for $\bar{S} \in \mathcal{A}$, since for other sets S, f_S equals $\perp$ and is at distance 1 from any other function.

Let $g' : [N]^k \to [M]^k$ be the direct product function that for every $i \in [k]$ has $g'_i : [N] \to [M]$ be the functions defined by $g(i)$. If there is i such that $g(i) = \perp$, we define g'_i arbitrarily. By Theorem 5.13, for every $f_S \in C$, $\bar{S} \in \mathcal{A}$ and $g_{\bar{S}}, g$ are 50β-consistent. ◻

We now prove Theorem 5.13. In order to prove the theorem, it is helpful to look at the sets $S \in \binom{[k]}{9k/10}$ as vertices in a graph. Let $\mathcal{G} = (V, E_S \cup E_W)$ to be the graph with the vertex set $V = \binom{[k]}{9k/10}$, and two edge sets, weak edges and strong edges.

Definition 5.15. For every two sets $S_1, S_2 \in V$,

1. S_1, S_2 are connected by a strong edge, denoted by $S_1 - S_2$, if $\Delta(f_{S_1}, f_{S_2}) < \beta$.
2. S_1, S_2 are connected by a weak edge, denoted by $S_1 \sim S_2$, if $\Delta(f_{S_1}, f_{S_2}) < 10\beta$.

If S_1, S_2 are not connected by a weak edge, we denote $S_1 \not\sim S_2$.

We want to find a very dense set of vertices in $\mathcal{G}$. Such a subset will allow us to define a global function g. We start by showing that there are many vertices with high degree in $\mathcal{G}$.

Claim 5.16. *There exists a set $\mathcal{S} \subset V$ of measure at least $\nu/2$, such that for every $S \in \mathcal{S}$*

$$\Pr_{S'\in V}[S - S'] \geq \frac{1}{2}\nu\,.$$

Proof. Let

$$\mathcal{S} = \left\{ S \subseteq V \,\middle|\, \Pr_{S'}[S - S'] \geq \frac{1}{2}\nu \right\}.$$

By averaging

$$\begin{aligned}
\nu &\leq \Pr_{S_1,S_2}[S_1 - S_2] \\
&\leq \Pr_{S_1}[S_1 \in \mathcal{S}] \Pr_{S_1,S_2}[S_1 - S_2 \mid S_1 \in \mathcal{S}] + \Pr_{S_1}[S_1 \notin \mathcal{S}] \Pr_{S_1,S_2}[S_1 - S_2 \mid S_1 \notin \mathcal{S}] \\
&\leq \Pr_{S_1}[S_1 \in \mathcal{S}] + \frac{1}{2}\nu\left(1 - \Pr_{S_1}[S_1 \in \mathcal{S}]\right).
\end{aligned}$$

Then $\Pr_{S_1}[S_1 \in \mathcal{S}] \geq \nu/2$. ◻

Strong connectivity is not transitive, but we can have an "almost transitive" property by considering both strong and weak edges.

Claim 5.17. *For $S, S_1, S_2 \in V$ uniformly and independently,*

$$\Pr_{S,S_1,S_2}[S - S_1, S - S_2, S_1 \not\sim S_2] \le 2\mathrm{e}^{-\beta^2 k}\,.$$

From the claim we get that if S is connected to S_1, S_2 by a strong edge, then almost always S_1, S_2 are connected by a weak edge.

Proof. Fix $S_1, S_2 \in V$ to be two vertices such that $S_1 \not\sim S_2$ (if there are no such vertices, the probability is 0 and we are done). For every $i \in [k]$ let $d_i \in [0,1]$ be

$$d_i = \begin{cases} \mathrm{dist}(f_{S_1}(i), f_{S_2}(i)) & \text{if } i \in S_1 \cap S_2 \\ 0 & \text{otherwise.} \end{cases}$$

Since $S_1 \not\sim S_2$, $\sum_{i\in[k]} d_i \ge (8k/10) \cdot 10\beta = 8\beta k$ (the minimal size of $S_1 \cap S_2$ is $8k/10$).

If S is a set that is strongly connected to both S_1 and S_2, then by the triangle inequality

$$\sum_{i\in S\cap S_1\cap S_2} \mathrm{dist}(f_{S_1}(i), f_{S_2}(i)) \le \sum_{i\in S\cap S_1\cap S_2} \mathrm{dist}(f_S(i), f_{S_1}(i)) + \mathrm{dist}(f_S(i), f_{S_2}(i)) \le 2\beta k\,.$$

That is, $\sum_{i\in S} d_i \le 2\beta k$.

The set S is a uniform subset of $[k]$ of size $9k/10$. Using the Hoeffding bound for random sampling without replacement (Fact 2.5)

$$\Pr_S[S - S_1 \text{ and } S - S_2] \le \Pr_S\left[\sum_{i\in S} d_i \le 2\beta k\right] \le \mathrm{e}^{-2\beta^2 k}\,.$$

Since the bound holds for every $S_1 \not\sim S_2$, then it holds also for the uniform distribution over sets S_1, S_2. □

From the last two claims, Claim 5.16 and Claim 5.17, we conclude that there is a high degree vertex in V whose neighbors form a very dense graph with respect to weak edges.

Claim 5.18. *There exists a set $S \in \mathcal{S}$ such that*

$$\Pr_{S_1,S_2\in\mathcal{V}}[S_1 \sim S_2 \mid S_1 - S, S_2 - S] \ge 1 - \beta\,.$$

Proof. From Claim 5.16, we know that if we choose $S, S_1, S_2 \in V$ independently,

$$\Pr_{S,S_1,S_2}[S \in \mathcal{S}, S - S_1, S - S_2] \ge \Pr_S[S \in \mathcal{S}] \Pr_{S,S_1}[S - S_1 \mid S \in \mathcal{S}]^2 \ge \left(\frac{\nu}{2}\right)^3\,.$$

From Claim 5.17, on the same distribution

$$\Pr_{S,S_1,S_2}[S - S_1, S - S_2, S_1 \not\sim S_2] \le \mathrm{e}^{-2\beta^2 k}.$$

Therefore

$$\Pr_{S,S_1,S_2}[S_1 \not\sim S_2 \mid S \in \mathcal{S}, S - S_1, S - S_2] \le \left(\frac{2}{\nu}\right)^3 \mathrm{e}^{-2\beta^2 k}.$$

Since $\nu > \mathrm{e}^{-\frac{1}{3}\beta^2 k}$, the bound on the probability $\left(\frac{2}{\nu}\right)^3 \mathrm{e}^{-2\beta^2 k}$ is tiny and surely smaller than β.

By averaging, there must be $S \in \mathcal{S}$ that achieves this bound. ◻

Claim 5.19. *Let $C \subset V$ satisfy $\Pr_{S\in V}[S \in C] \ge \nu/2$, then the number of indices $i \in k$ such that $\Pr_{S\in C}[i \in S] \le 1/2$ is at most βk.*

Proof. Let $D \subset [k]$ be the set

$$D = \left\{ i \in [k] \,\middle|\, \Pr_{S\in C}[i \in S] \le \frac{1}{2} \right\}.$$

If we pick a completely uniform $S \in \binom{[k]}{\frac{9}{10}k}$, then $\mathbb{E}_S[|S \cap D|] = \frac{9}{10}|D|$.

Using the the Hoeffding bound for random subset (Fact 2.4), $\Pr_S\left[|S \cap D| \le \frac{2}{3}|D|\right] \le \mathrm{e}^{-\frac{|D|}{100}}$. If instead we pick a uniform subset in $S \in C$, the probability of the event $|S \cap D| \le \frac{2}{3}|D|$ may increase by a factor of $2/\nu$.

$$\Pr_{S\in C}\left[|S \cap D| \le \frac{2}{3}|D|\right] \le \frac{2}{\nu}\mathrm{e}^{-\frac{|D|}{100}}.$$

By the definition of D, for each $i \in D$, $\Pr_{S\in C}[i \in S] \le \frac{1}{2}$, so $\mathbb{E}_{S\in C}[|S \cap D|] \le |D|/$. From averaging $\Pr_{S\in C}[|S \cap D| \le 2|D|/3] \ge 1/4$.

Combining the two bounds, we get that $2/\nu \cdot \mathrm{e}^{-\frac{|D|}{100}} \ge 1/4$, which means that $|D| \le \beta k$ (recall that $\nu > \mathrm{e}^{-\beta^2 k}$ and β is a small constant). ◻

Proof of Theorem 5.5. Let $\tilde{S} \in \mathcal{S}$ be the vertex promised from Claim 5.18, and denote by C its strong neighbors,

$$C = \left\{ S \in V \,\middle|\, S - \tilde{S} \right\}.$$

The measure of C is at least $\frac{\nu}{2}$ and $\Pr_{S_1,S_2\in C}[S_1 \not\sim S_2] \le \beta$, which implies that

$$\mathop{\mathbb{E}}_{S_1,S_2\in C}[\Delta(f_{S_1}, f_{S_2})] \le 1 \cdot \Pr_{S_1,S_2\in C}[S_1 \not\sim S_2] + \mathop{\mathbb{E}}_{S_1,S_2\in C}[\Delta(f_{S_1}, f_{S_2}) \mid S_1 \sim S_2] \le \beta + 10\beta. \tag{5.2}$$

We define our direct product function $g : [k] \to \Sigma$ as follows.

$$\forall i \in [k] \quad g(i) = \arg\min_{\sigma\in\Sigma} \left\{ \mathop{\mathbb{E}}_{S\in C \text{ s.t. } i\in S}[\mathrm{dist}(f_S(i), \sigma)] \right\}.$$

Ties are broken arbitrarily. If there is no $S \in C$ such that $i \in S$, we set $g(i)$ to an arbitrary value.

We prove that the function g satisfies the theorem requirements. It is enough to prove that $\Delta(g, f_S) < 50\beta$ for half of the sets $S \in C$.

We want to bound the expected difference $\mathbb{E}_{S\in C}[\Delta(g, f_S)]$. If we could say that $\mathbb{E}_{S\in C}[\Delta(g, f_S)] \leq \mathbb{E}_{S_1,S_2\in C}[\Delta(f_{S_1}, f_{S_2})]$, then by (5.2) we would be done. Unfortunately there is a slight complication, which we explain and solve below. We write the two expectations explicitly.

1. $\mathbb{E}_{S\in C}[\Delta(g, f_S)] = \mathbb{E}_{S\in C, i\in S}[\mathrm{dist}(g(i), f_S(i)]$. Let $\mathcal{D}_1 : [k] \to [0, 1]$ be the distribution over i in this expectation, i. e., picking $S \in C$ and then a uniform $i \in S$.

2. $\mathbb{E}_{S_1,S_2\in C}[\Delta(f_{S_1}, f_{S_2})] = \mathbb{E}_{S_1,S_2\in C, i\in S_1\cap S_2}[\mathrm{dist}(f_{S_1}(i), f_{S_2}(i))]$. Let $\mathcal{D}_2 : [k] \to [0, 1]$ be the distribution over i described in this expectation, i. e., picking $S_1, S_2 \in C$ and then $i \in S_1 \cap S_2$.

We prove that the two distributions are rather similar. Let D be the set of "bad locations", which appear too little in C,

$$D = \left\{ i \in [k] \,\middle|\, \Pr_{S\in C}[i \in S] < \frac{1}{2} \right\} .$$

By Claim 5.19, $|D| \leq \beta k$, and clearly also $\Pr_{i\sim\mathcal{D}_1}[i \in D] \leq \beta$.

For every $i \notin D$, the index i appears in at least half of the sets $S \in C$. This means that for every such i, $\mathcal{D}_2(i) \geq \mathcal{D}_1(i)/2$.

For every $i \in [k]$, by the definition of g,

$$\mathop{\mathbb{E}}_{S\in C}[\mathrm{dist}(f_S(i), g(i))|i \in S] \leq \mathop{\mathbb{E}}_{S_1,S_2\in C}[\mathrm{dist}(f_{S_1}(i), f_{S_2}(i))|i \in S_1 \cap S_2] . \tag{5.3}$$

Therefore,

$$\begin{aligned}
\mathop{\mathbb{E}}_{S\in C}[\Delta(g, f_S)] &= \mathop{\mathbb{E}}_{i\sim\mathcal{D}_1, S\in C}[\mathrm{dist}(g(i), f_S(i)|i \in S] \\
&\leq \Pr_{i\sim\mathcal{D}_1}[i \in D] + \mathop{\mathbb{E}}_{i\sim\mathcal{D}_1, S\in C}[\mathrm{dist}(g(i), f_S(i))|i \notin D, i \in S] \\
&\leq \beta + \mathop{\mathbb{E}}_{i\sim\mathcal{D}_1, S_1,S_2\in C}[\mathrm{dist}(f_{S_1}(i), f_{S_2}(i))|i \notin D, i \in S_1 \cap S_2] && \text{(by (5.3))} \\
&\leq \beta + 2 \mathop{\mathbb{E}}_{i\sim\mathcal{D}_2, S_1,S_2\in C}[\mathrm{dist}(f_{S_1}(i), f_{S_2}(i))|i \notin D, i \in S_1 \cap S_2] \\
&\leq \beta + 2 \cdot 11\beta \leq 25\beta . && \text{(by (5.2))}
\end{aligned}$$

To finish the proof, the only thing left is a Markov argument. If $\mathbb{E}_{S\in C}[\Delta(f_S, g)] \leq 25\beta$, then at least half of the sets $S \in C$ satisfies $\Delta(f_S, g) \leq 50\beta$, and we finish the proof. □

5.4 Global direct product function

Proof of Lemma 5.6. Recall $\mathcal{A}$ is the set of sets which are successful with respect to our fixed string x. From Lemma 5.4,

$$\Pr_{A_1,A_2\in\binom{[k]}{k/10}}[A_1, A_2 \in \mathcal{A} \text{ and } g_{A_1}, g_{A_2} \text{ are } \lambda_1\text{-consistent}] \geq \epsilon^5.$$

We apply Theorem 5.5 on the family of functions $\mathcal{F} = \{g_A\}_{A \in \mathcal{A}}$. Let $g : [N]^k \to [M]^k$ be the global direct product function promised from the theorem. Fix $\lambda_2 = 50\lambda_1$ and let $\mathcal{A}^*$ be

$$\mathcal{A}^* = \{A \mid \ A \in \mathcal{A},\ g_A \text{ is } \lambda_2\text{-consistent with } g\}\,.$$

From the theorem, $\Pr_{A \in \binom{[k]}{k/10}}[A \in \mathcal{A}^*] \geq \epsilon^5/4$.

For every $z \in [N]^k, A \in \mathcal{A}^*$, let $E(A,z)$ be the event

$$f(z)_{\bar{A}} \overset{\lambda_1 k}{\approx} g_A(z_{\bar{A}}) \quad \text{and} \quad g_A(z_{\bar{A}}) \overset{2\lambda_2 k}{\approx} g(z)_{\bar{A}}\,.$$

For every $A \in \mathcal{A}^*$, the direct product functions g_A, g are λ_2-consistent. By the Chernoff bound (Fact 2.3), for a uniform input $z \in [N]^k$,

$$\Pr_z[g_A(z_{\bar{A}}) \overset{2\lambda_2 k}{\not\approx} g(z)_{\bar{A}}] \leq \mathrm{e}^{-\frac{\lambda_2 k}{3}}\,.$$

From Claim 5.8, for every $A \in \mathcal{A}$, $\Pr_z[f(z)_{\bar{A}} \overset{\lambda_1 k}{\approx} g_A(z_{\bar{A}})] \geq \epsilon/4$.

Therefore, for every $A \in \mathcal{A}^*$,

$$\Pr_{z \in [N]^k}[E(A,z)] \geq \Pr_z[f(z)_{\bar{A}} \overset{\lambda_1 k}{\approx} g_A(z_{\bar{A}})] - \Pr_z[g_A(z) \overset{2\lambda_2 k}{\not\approx} g(z)_{\bar{A}}] \geq \frac{\epsilon}{4} - \mathrm{e}^{-\frac{\lambda_2 k}{3}}\,.$$

The same bound holds also for a uniform $A \in \mathcal{A}^*$.

Let $\mathcal{Z}$ be the set of inputs,

$$\mathcal{Z} = \left\{ z \in [N]^k \;\middle|\; \Pr_{A \in \mathcal{A}^*}[E(A,z)] \geq \frac{\epsilon}{8} \right\}.$$

From averaging,

$$\Pr_{z \in [N]^k, A \in \mathcal{A}^*}[E(A,z)] \leq \Pr_z[z \in \mathcal{Z}] \cdot 1 + \Pr_z[z \notin \mathcal{Z}]\frac{\epsilon}{8}\,.$$

That is, $\Pr_z[z \in \mathcal{Z}] \geq \epsilon/4 - \mathrm{e}^{-\frac{\lambda_2 k}{3}} - \epsilon/8 \geq \epsilon/10$. Fix $\lambda_3 = 3/2\,(\lambda_1 + 2\lambda_2)$. We prove that for every $z \in \mathcal{Z}$, $f(z) \overset{\lambda_3 k}{\approx} g(z)$, which finishes the proof.

Fix $z \in \mathcal{Z}$ and let $D \subset [k]$ be the set

$$D = \{i \in [k] \mid \ f(z)_i \neq g(z)_i\}\,.$$

Assume for a contradiction that $f(z) \overset{\lambda_3 k}{\not\approx} g(z)$, i. e., $|D| > \lambda_3 k$.

For each $A \in \mathcal{A}^*$ such that $E(A,z)$ happen, by the triangle inequality, $f(z)_{\bar{A}} \overset{\lambda_1 k + 2\lambda_2 k}{\approx} g(z)_{\bar{A}}$. This can only happen if A is such that $|\bar{A} \cap D| \leq \lambda_1 k + 2\lambda_2 k = 2|D|/3$. The set A is a uniform subset of $[k]$ of size $k/10$. By the Hoeffding bound for random subset (Fact 2.4), $\Pr_A[|\bar{A} \cap D| \leq 2|D|/3] \leq \mathrm{e}^{-\frac{k}{20}\lambda_3}$. We conclude that $\Pr_A[A \in \mathcal{A}^* \text{ and } E(A,z)] \leq \mathrm{e}^{-\frac{k}{20}\lambda_3}$.

This contradicts $z \in \mathcal{Z}$, because for $z \in \mathcal{Z}$,

$$\Pr_A[A \in \mathcal{A}^* \text{ and } E(A,z)] \geq \Pr_A[A \in \mathcal{A}^*] \Pr_A[E(A,z) | A \in \mathcal{A}^*] \geq \frac{1}{4}\epsilon^5 \cdot \frac{\epsilon}{8}\,. \qquad \square$$

6 Lower bounds for approximate equality

In this section we prove lower bounds for different variants of the direct product test. The lower bounds are proven by finding a function that passes the test with some probability ϵ, but is far from any direct product function.

We start by defining when a function is far from any direct product function.

Definition 6.1. Two functions $f_1, f_2 : [N]^k \to [M]^k$ are (ϵ, λ) close, if

$$\Pr_{x \in [N]^k}\left[f_1(x) \overset{\lambda k}{\approx} f_2(x)\right] \geq \epsilon\,.$$

A function $f : [N]^k \to [M]^k$ is (ϵ, λ) close to a direct product function if there are functions $g_1, \ldots, g_k : [N] \to [M]$ such that $(g_1, \ldots, g_k)$ is (ϵ, λ) close to f. Otherwise, $f : [N]^k \to [M]^k$ is (ϵ, λ) far from any direct product function.

Our direct product theorem states that if a function $f : [N]^k \to [M]^k$ passes Test 1 with $t = k/10$ with probability $\epsilon > \mathrm{e}^{-c\lambda^2 k}$, then f is $(\epsilon/10, \lambda)$ close to a direct product function. Ideally, we want the even stronger conclusion that f is $(\Omega(\epsilon), 0)$ close to a direct product function. However, we next show that this is not possible.

Claim 6.2. *For every $N \in \mathbb{N}$ and large enough k, the following holds. Let $f : [N]^k \to \{0,1\}^k$ be a random function, i. e., such that $f(x)$ is a uniformly chosen string in $[M]^k$ for each x independently. With high probability, f is $(2\mathrm{e}^{-k/10}, k/10)$ far from any direct product function.*

Proof. Fix a direct product function $g : [N]^k \to \{0,1\}^k$. For every $x \in [N]^k$, by the Chernoff bound, $\Pr_f[f(x) \overset{k/10}{\approx} g(x)] \leq \mathrm{e}^{-k/10}$. The probability that $f(x) \overset{k/10}{\approx} g(x)$ on more than $2\mathrm{e}^{-k/10}$ of the inputs $x \in [N]^k$ is smaller than $\mathrm{e}^{-\frac{1}{3}N^k \cdot \mathrm{e}^{-k/10}} \leq \mathrm{e}^{-\frac{1}{3}\left(\frac{N}{2}\right)^k}$.

There are 2^{Nk} different direct product functions $g : [N]^k \to \{0,1\}^k$. By union bound, with probability at least $1 - 2^{Nk}\mathrm{e}^{-\frac{1}{3}\left(\frac{N}{2}\right)^k}$, the random function f is $(2\mathrm{e}^{-k/10}, k/10)$ far from any direct product function. $\square$

6.1 Testing different intersection sizes.

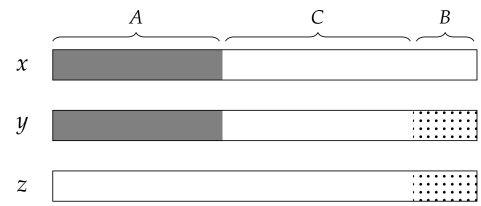
We next analyze a family of tests, parameterized by $t_1, t_2 \in [k]$, that generalize our basic Z-test and appear as Test 5. We show that there is no direct product testing theorem for these tests, with $(\Omega(\epsilon), 0)$ closeness, for an exponentially small ϵ.

Claim 6.3. *For every constant $\delta > 0$, large enough $k, N, M \in \mathbb{N}$ such that $N \geq M \geq k^{12}$ and every $t_1, t_2 \in [k]$ such that $t_1 + t_2 \leq k$, there exists a constant $\beta > 0$ and a function $f : [N]^k \to [M]^k$, such that $agr^Z_{t_1,t_2}(f) = \epsilon \geq \mathrm{e}^{-\delta k}$, but f is $(\epsilon^2, \frac{\beta}{\log k})$ far from any direct product function.*

Proof. Let β be a constant that will be determined later, and denote $\ell = \frac{2\beta k}{\log k}$. Let $h : [N] \to [M] \setminus \{1\}$ be any function satisfying: for every $a \in [M] \setminus \{1\}$, $\Pr_{b \in [N]}[h(b) = a] \leq 2/M$.

Test 5: Z-test with parameters t_1, t_2 (3-query test)

1. Choose A, B, C to be a random partition of $[k]$, such that $|A| = t_1, |B| = t_2$.
2. Choose uniformly at random $x, y, z \in [N]^k$ such that $x_A = y_A$ and $y_B = z_B$.
3. Accept if $f(x)_A = f(y)_A$ and $f(z)_B = f(y)_B$.



Denote by $agr^Z_{t_1,t_2}(f)$ the success probability of f on this test.

Let $f : [N]^k \to [M]^k$ be the following function: for every $x \in [N]^k$ let $i_1, \ldots, i_\ell, j_1, \ldots, j_\ell \in [k]$ be different random coordinates.

$$\forall i \in [k], \quad f(x)_i = \begin{cases} h(x_{j_r}) & \text{if } i = i_r \text{ for some } r \in [\ell] \\ 1 & \text{otherwise.} \end{cases}$$

The function f has ℓ random "corrupted" coordinates per input, $i_1, \ldots, i_\ell$, in which $f(x)_{i_r}$ is determined by the value of x_{j_r}.

We analyze the success probability of Test 5 with parameters t_1, t_2 on f. We divide into two cases.

1. In case $\max\{t_1, t_2\} \leq 0.4k$. Let x, y, z, A, B be the sets and strings chosen by the test. Then $|A \cup B| \leq 0.8k$. Let $i^x_1, \ldots, i^x_\ell$ be the corrupted coordinates of x (and i^y_r, i^z_r for y, z). If $i^x_1, \ldots, i^x_\ell, i^y_1, \ldots, i^y_\ell, i^z_1, \ldots, i^z_\ell \notin A \cup B$, the test "misses" all of the corrupted coordinates, so the test passes.

   $$\Pr_f[i^x_1, \ldots i^x_\ell \notin A \cup B] = \Pr_f[i^x_1 \notin A \cup B] \cdots \Pr_f[i^x_\ell \notin A \cup B | i^x_1 \ldots, i^x_{\ell-1} \notin A \cup B] \geq (0.1)^\ell,$$

   where the last inequality is because $|A \cup B| \leq 0.8k$ and $\ell < 0.1k$. Therefore, even conditioning on $i^x_1 \ldots, i^x_{\ell-1} \notin A \cup B$, the probability of $i^x_\ell \notin A \cup B$ is at least 0.1. The same inequality holds also for y and z, and we get that f passes Test 5 with probability at least $(0.1)^{3\ell}$.

2. In case $\max\{t_1, t_2\} > 0.4k$. The test is symmetric with respect to t_1, t_2, so we can assume w.l.o.g. that $t_1 \geq t_2$. Let x, y, z, A, B be the sets and strings chosen by the test. Let $i^x_1, \ldots, i^x_\ell$ and $j^x_1, \ldots, j^x_\ell$ be the chosen coordinates of x, and the same for y and z. If for all $r \in [\ell]$, $i^x_r = i^y_r$ and $j^x_r = j^y_r$ and also $i^x_1, \ldots, i^x_\ell \in A$ and $j^x_1, \ldots, j^x_r \in A$, then the corrupted coordinates of x and y are corrupted to the same value. If in addition $i^z_1, \ldots, i^z_\ell \in A$, then the corrupted coordinates of z are not checked, and the test passes. We lower bound the probability of these events.

   $$\Pr_f[\forall r \in [\ell], i^x_r = i^y_r \text{ and } j^x_r = j^y_r] = \frac{1}{k} \cdot \frac{1}{k-1} \cdots \frac{1}{k-2\ell+1} \geq \left(\frac{1}{k}\right)^{2\ell}.$$

This is because the probability of $i_1^x = i_1^y$ is $\frac{1}{k}$, and the probability of $i_2^x = i_2^y$ given that $i_1^x = i_1^y$ is $1/(k-1)$, and so forth.

$$\Pr_f[i_1^x, \dots, i_\ell^x \in A \text{ and } j_1^x, \dots j_\ell^x \in A]$$

$$= \Pr_f[i_1^x \in A] \cdots \Pr_f[j_\ell^x \in A | i_1^x, \dots, i_\ell^x, j_1^x, \dots, j_{\ell-1}^x \in A] \geq (0.3)^{2\ell},$$

where the last inequality is because $|A| \geq 0.4k$, and $\ell < 0.05k$. Therefore, f passes the test with probability at least $(1/k)^{2\ell} \cdot (0.3)^{2\ell} \cdot (0.3)^{2\ell}$. We pick the constant β such that $\ell = 2\beta k/(\log k)$ satisfies $(1/k)^{2\ell} \cdot (0.3)^{4\ell} \geq \mathrm{e}^{-\delta k}$.

We prove next that f is $(\epsilon^2, \beta/(\log k))$ far from any direct product function. For every $S \subset [k], |S| = \frac{\ell}{2}$, $w \in [N]^S$ and $\gamma \in ([M] \setminus \{1\})^S$, let $\mathcal{S}_{S,w,\gamma} = \left\{ x \in [N]^k \;\middle|\; x_S = w, f(x)_S = \gamma \right\}$.

We say that f is balanced if for every S, w, γ,

$$\Pr_x[x \in \mathcal{S}_{S,w,\gamma} | x_S = w] \leq \left(\frac{2k}{M}\right)^{\frac{\ell}{2}}.$$

We prove that f is balanced. Fix any S, w and γ. For every x, $f(x)_S = \gamma$ only if for every $i \in S$, $f(x)_i$ is chosen to be corrupted to some $h(x_j) = \gamma_i$. If there is no j such that $h(x_j) = \gamma_i$, it is not possible that $f(x)_S = \gamma$. For each γ_i, the probability over $x \in [N]^k, x_S = w$ that $x_{\bar{S}}$ contains a coordinate j such that $h(x_j) = \gamma_i$ is smaller than $2k/M$. Therefore, the probability over x that there are $\frac{\ell}{2}$ different coordinates $j_1, \dots, j_{\frac{\ell}{2}} \in \bar{S}$ such that for all i, $h(x_{j_i}) = \gamma_i$ is less than $(2k/M)^{\frac{\ell}{2}}$ (if $\gamma_i = \gamma_r$, then x needs to contain two different coordinates j_i, j_r such that $h(x_{j_i}) = h(x_{j_r}) = \gamma_i$). From this we deduce that for every $S \subset [k], |S| = \frac{\ell}{2}, w \in [N]^S$, $\Pr_x[x \in \mathcal{S}_{S,w,\gamma} | x_S = w] \leq (2k/M)^{\frac{\ell}{2}}$, and f is balanced.

We prove that a balanced function f is far from any direct product function. Fix a direct product function $g : [N]^k \to [M]^k$, and let $F = \left\{ x \in [N]^k \;\middle|\; f(x) \overset{\ell/2}{\approx} g(x) \right\}$. For every $x \in F$, $f(x)$ and $g(x)$ are equal on at least $\ell/2$ "corrupted" coordinates, i. e., for every $x \in F$ there exists a set $S \subset [k], |S| = \ell/2$ such that $f(x)_S = g(x)_S \in ([M] \setminus \{1\})^S$.

For every $S \subset [k], |S| = \frac{\ell}{2}$ and $w \in [N]^S$ let

$$F_{S,w} = \left\{ x \in F \;\middle|\; x_S = w, f(x)_S = g(x)_S \in ([M] \setminus \{1\})^S \right\}.$$

From above, $F \subset \cup_{S,w} F_{S,w}$. The function g is a direct product function, so for every S and w there is a single γ such that $g(x)_S = \gamma$ for *all* x such that $x_S = w$. That is, $F_{S,w} \subset \mathcal{S}_{S,w,\gamma}$ for some $\gamma \in ([M] \setminus \{1\})^S$. Together we get that $|F| \leq \sum_{S,w} |F_{S,w}| \leq \binom{k}{\frac{\ell}{2}} \cdot (2k/M)^{\ell/2} N^k$. Since $M, N \geq k^{12}$ we get that $\Pr_x[x \in F] \leq \binom{k}{\frac{\ell}{2}} (2k/M)^{\ell/2} \leq (2/k)^{5\ell} \leq \epsilon^2$. □

6.2 The triangle test for functions over sets.

In Test 5 with $t_1 + t_2 = k$, we check $f(y)$ on all coordinates, but only part of the coordinates of $f(x), f(z)$. What if we check all coordinates of all three inputs? This brings us to the triangle test, Test 6, for functions over sets. In this test, every two out of the three inputs in the test share a joint subset of size $\frac{k}{2}$. For this test we must assume that k is even.

We remark that Test 5 can only be defined on a function $f : \binom{[N]}{k} \to [M]^k$, and it is not possible to define such test on a function on tuples. This is because for $f : [N]^k \to [M]^k$, if we choose inputs x, y, z such that $x_A = y_A, y_B = z_B$, it is not possible to compare $f(x)_B$ to $f(z)_A$, since these are different coordinates. We remark that it is possible to define a 4-query test on functions over tuples, similar to the triangle test, with inputs x, y, z, w such that $x_A = y_A, z_A = w_A, x_B = z_B, y_B = w_B$, but we do not analyze it here.

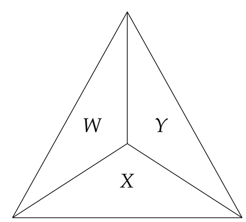
We define distance from a direct product function in a similar way to functions over tuples.

Definition 6.4. A function $f : \binom{[N]}{k} \to [M]^k$ is (ϵ, λ) far from any direct product function, if for every $g : [N] \to [M]$,

$$\Pr_{S \in \binom{[N]}{k}} [f(S) \overset{\lambda k}{\approx} g(S)] \leq \epsilon.$$

Test 6: Triangle test (3-query test, for even k)

1. Choose disjoint $W, X, Y \subset [N]$ of size $\frac{k}{2}$.
2. Accept if $f(X \cup W)_W = f(Y \cup W)_W$, $f(X \cup Y)_Y = f(Y \cup W)_Y$ and $f(X \cup W)_X = (X \cup Y)_X$.



Denote by $agr^{\Delta}(f)$ the success probability of f on this test.

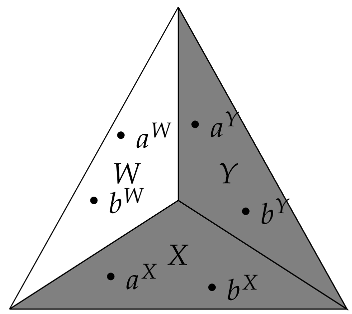
Claim 6.5. *For every constant $\delta > 0$, large enough $k, N, M \in \mathbb{N}$ such that $N \geq M \geq k^{30}$ there exists a constant $\beta > 0$ and a function $f : \binom{[N]}{k} \to [M]^k$, such that $agr^{\Delta}(f) = \epsilon > \mathrm{e}^{-\delta k}$, and f is $(\epsilon^2, \frac{\beta}{\log k})$ far from any direct product function.*

Proof. We prove the claim by constructing a function f that passes the test but is far from any direct product function.

Let $\beta > 0$ be a constant that will be decided later, and let $\ell = \beta k/(\log k)$. We describe a function $f : \binom{[N]}{k} \to [M]^k$ with 2ℓ "corrupted" elements per input. Let $h : [N] \to [M] \setminus \{1\}$ be an arbitrary function such that for every $\gamma \in [M] \setminus \{1\}$, $\Pr_{b \in [N]}[h(b) = \gamma] \leq 2/M$. For every $S \in \binom{[N]}{k}$ we pick 2ℓ elements to corrupt f on $a_1, \ldots, a_{2\ell}$ and 2ℓ index elements $b_1, \ldots, b_{2\ell} \in S$. The function f is defined by:

$$\forall e \in S, \quad f(S)_e = \begin{cases} h(b_i) & \text{if } e = a_i \text{ for some } i \in [2\ell] \\ 1 & \text{otherwise.} \end{cases}$$

Figure 2: The set $X \cup Y$ is marked in gray.



We analyze the success probability of the test. Let W, X, Y be the sets chosen by the test algorithm. Let $a_1^X, \ldots, a_\ell^X, b_1^X, \ldots, b_\ell^X \in X$ be arbitrary 2ℓ elements in X, and the same for Y, W. If f is such that

1. In $f(W \cup X)$, the corrupted elements are $a_1^X, \ldots, a_\ell^X$ and $a_1^W, \ldots, a_\ell^W$, and the index elements are $b_1^X, \ldots, b_\ell^X$ and $b_1^W, \ldots, b_\ell^W$.
2. In $f(X \cup Y)$, the corrupted elements are $a_1^X, \ldots, a_\ell^X$ and $a_1^Y, \ldots, a_\ell^Y$, and the index elements are $b_1^X, \ldots, b_\ell^X$ and $b_1^Y, \ldots, b_\ell^Y$.
3. In $f(W \cup Y)$, the corrupted elements are $a_1^W, \ldots, a_\ell^W$ and $a_1^Y, \ldots, a_\ell^Y$, and the index elements are $b_1^W, \ldots, b_\ell^W$ and $b_1^Y, \ldots, b_\ell^Y$.

Then the corrupted elements are corrupted to the same value on all three queries $f(X \cup W)$, $f(X \cup Y)$ and $f(Y \cup W)$ and the test passes. See Figure 2 for an illustration (with $\ell = 1$).

The probability that in $f(X \cup W)$, the 2ℓ corrupted elements are $a_1^X, \ldots, a_\ell^X$ and $a_1^W, \ldots, a_\ell^W$ and the index elements are $b_1^X, \ldots, b_\ell^X$ and $b_1^W, \ldots, b_\ell^W$ is larger than $\frac{1}{k^{4\ell}}$. Therefore, f satisfies $\mathrm{agr}^\Delta(f) \geq \frac{1}{k^{12\ell}}$. We choose the constant β to be small enough such that for $\ell = \frac{\beta k}{\log k}$, $\mathrm{agr}^\Delta(f) \geq \frac{1}{k^{12\ell}} \geq \mathrm{e}^{-\delta k}$.

We prove that f is $(\epsilon^2, \frac{\beta}{\log k})$-far from any direct product function. For every $L \subset [N], |L| = \ell$ and $\gamma \in ([M] \setminus \{1\})^L$, let $\mathcal{S}_{L,\gamma} = \left\{ S \in \binom{[N]}{k} \,\middle|\, L \subset S, f(S)_L = \gamma \right\}$.

We say that the function $f : \binom{[N]}{k} \to [M]^k$ is *balanced* if for every $L \subset [N], |L| = \ell$ and $\gamma \in ([M] \setminus \{1\})^L$,

$$\Pr_S[S \in \mathcal{S}_{L,\gamma} | L \subset S] \leq \left(\frac{4k}{M} \right)^\ell .$$

We claim that f is balanced. Fix L and γ, a set $S \in \mathcal{S}_{L,\gamma}$ if for every $a \in L$, $f(S)_a$ is corrupted to $h(b) = \gamma_a$, if S doesn't contain b such that $h(b) = \gamma_a$ then its not possible that $f(S)_a = \gamma_a$. For every γ_a, the probability of $S \subset [N], L \subset S$ to be such that $S \setminus L$ contains a coordinate b satisfying $h(b) = \gamma_a$ is at most $\frac{2k}{M} \frac{N}{N-\ell}$ (the requirement that $b \notin L$ can increase the probability of $h(b) = \gamma_a$

by a factor of at most $\frac{N}{N-\ell} < 2$). The probability that $S \setminus L$ contains ℓ elements $b_1, \ldots, b_\ell$ such that for all a there is a different b_a satisfying $h(b_a) = \gamma_a$ is at most $\left(\frac{2k}{M} \cdot \frac{N}{N-2\ell}\right)^\ell \leq \left(\frac{4k}{M}\right)^\ell$, so the function f is balanced (the factor $\frac{N}{N-2\ell}$ is because we require that the $b_i \notin L$, and are different).

We prove that a balanced f is far from every direct product function. Fix a direct product function $g : [N] \to [M]$, and let F be the set $F = \left\{ S \in \binom{[N]}{k} \,\middle|\, f(S) \stackrel{\ell}{\approx} g(S) \right\}$. For every $S \in F$, there is a set $L \subset S$ of size ℓ such that for every $a \in L$, $f(S)_a = g(a) \neq 1$. Therefore, $F \subset \cup_{L \in \binom{[N]}{\ell}} \mathcal{S}_{L,g(L)}$, and $g(L)$ is different from 1 on all elements. This implies an upper bound on $|F|$,

$$|F| \leq \sum_{L \in \binom{[N]}{\ell}} |\mathcal{S}_{L,g(L)}| \leq \binom{N}{\ell} \left(\frac{4k}{M}\right)^\ell \binom{N-\ell}{k-\ell} \leq \left(\frac{4k}{M}\right)^\ell \binom{N}{k}.$$

For our choice of parameters $N, M \geq k^{30}$ and f is $(\epsilon^2, \frac{\ell}{k})$ far from any direct product function. □

A Tuples to sets restricted global structure proof

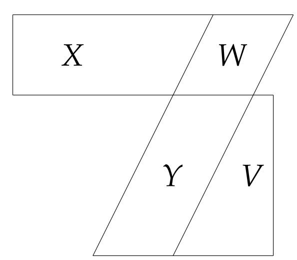
In this section we prove Lemma 4.4, restricted global structure for sets, which we restate below.

Lemma A.1 (Lemma 4.4, restated). *There exists a small constant $c > 0$, such that for every constant $\alpha > 0$, large enough $k \in \mathbb{N}$ and $N > \mathrm{e}^{c\lambda k}$, $M \in \mathbb{N}$, the following holds.*

For every function $f : \binom{[N]}{k} \to [M]^k$, if $agr^{Z_{set}}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\alpha k}$, then at least $(1 - \epsilon^2 - k^2/N)$ of the good pairs W, X are α-DP pairs.

The restricted global structure only uses the first two queries of the test. For convenience, we rewrite the test such that the two checks are not preformed in the same step.

Test 7: Z-test for functions over sets, with $t = k/10$ (3-query test)



1. Choose a random set $W \subset [N]$ of size $k/10$.
2. Choose $X, Y \subset [N] \setminus W$ of size $9k/10$.
3. If $f(X \cup W)_W \neq f(Y \cup W)_W$ reject.
4. Choose $V \subset [N] \setminus Y$ of size $k/10$.
5. If $f(Y \cup W)_Y \neq f(Y \cup V)_Y$ reject, else accept.

Denote by $agr^{Z_{set}}_{k/10}(f)$ the success probability of f on this test.

We prove the lemma by a reduction from Theorem 3.9.

Definition A.2. We associate each $S \subset [N]$ with the tuple $\vec{S} \in [N]^{|S|}$ obtained by sorting the elements of S in increasing order. For every string $x \in [N]^k$, we define $U(x) = 1$ if x has distinct coordinates, i. e., there is no $i \neq j$ such that $x_i = x_j$, else $U(x) = 0$.

For a string $s = (s_1, \dots, s_k)$ and a permutation $\pi : [k] \to [k]$ let $s^\pi = (s_{\pi(1)}, \dots, s_{\pi(k)})$.

Definition A.3. Given a function $f : \binom{[N]}{k} \to [M]^k$, let $f' : [N]^k \to [M]^k \cup \perp$ be defined as follows. If $U(x) = 0$ we set $f'(x) = \perp$. If $U(x) = 1$ then if $x = \vec{X}$ (namely $x_1 < x_2 < \cdots < x_k$) we let $f(x) = f(X)$. Otherwise there is some permutation π such that $x = (\vec{X})^\pi$ and we let $f'(x) = f(\vec{X})^\pi$. We call π the *sorting permutation* for the tuple x.

When testing the function f', whenever the tester queries an input x such that $f'(x) = \perp$, we assume the tester rejects.

Definition A.4. Let $\mathcal{D} : \binom{[N]}{k/10} \times \binom{[N]}{9k/10} \times \binom{[N]}{9k/10} \to [0,1]$ be the following distribution:

1. Choose $W \subset [N]$ of size $k/10$.
2. Choose $X \subset [N]$ of size $9k/10$ such that $X \cap W = \emptyset$.
3. Choose $Y \subset [N]$ of size $9k/10$ such that $Y \cap W = \emptyset$.

Let $\mathcal{D}' : \binom{[k]}{k/10} \times [N]^k \times [N]^k \to [0,1]$ be the following distribution:

1. Choose a set $A \subset [k]$ of size $k/10$.
2. Choose $x \in [N]^k$ such that $U(x) = 1$.
3. Choose $y \in [N]^k$ such that $x_A = y_A$ and $U(y) = 1$.

The distribution $(W, X, Y) \sim \mathcal{D}$ is the distribution used in Test 4. The distribution $(A, x, y) \sim \mathcal{D}'$ is the distribution of Test 2, conditioning on $U(x) = U(y) = 1$.

If we pick $(W, X, Y) \sim \mathcal{D}$, a random set $A \in \binom{[k]}{k/10}$ and random permutations $\pi_1 \in \mathcal{S}_{k/10}, \pi_2, \pi_3 \in \mathcal{S}_{9k/10}$, then $x = ((\vec{W}^{\pi_1})_A, (\vec{X}^{\pi_2})_{\bar{A}})$ and $y = ((\vec{W}^{\pi_1})_A, (\vec{Y}^{\pi_3})_{\bar{A}})$, are distributed according to $\mathcal{D}'$.

Let $p_1 = \Pr_{x \in [N]^k}[U(x) = 0]$. We bound its value. Choosing a uniform $x \in [N]^k$ can be done coordinate by coordinate. For each coordinate i, the probability that $x_i = x_j$ for some $j < i$ is at most $\frac{i-1}{N}$, therefore

$$p_1 = \Pr_{x \in [N]^k}[U(x) = 0] \leq \sum_{i=1}^{k} \frac{i-1}{N} \leq \frac{k^2}{2N}.$$

Fix an arbitrary $x \in [N]^k$ such that $U(x) = 1$ and a set $A \subset [k]$. Let

$$p_2 = \Pr_{y \in [N]^k}[U(y) = 0 \mid y_A = x_A].$$

We note that the value of p_2 does not depend on x, A. We can think of y as being chosen by starting from y_A, and choosing the rest of the coordinates one by one.

$$p_2 = \Pr_{y}[U(y) = 0 \mid y_A = x_A] \leq \sum_{i=k/10}^{k} \frac{i-1}{N} \leq \frac{k^2}{2N}.$$

We prove two claims connecting the test success probability on $f : \binom{[N]}{k} \to [M]^k$ to the success probability of the two-query test on the function $f' : [N]^k \to [M]^k \cup \bot$. Recall the notation of $\mathrm{agr}^{\vee}_{k/10}(\cdot)$ from Test 2.

Claim A.5. *For every function $f : \binom{[N]}{k} \to [M]^k$, the function $f' : [N]^k \to [M]^k \cup \bot$ from Definition A.3 satisfies*

$$agr^{\vee}_{k/10}(f') = (1 - p_1)(1 - p_2) \Pr[f \textit{ passes Item 3 of Test 4}].$$

Proof. Choose x, y, A according to the distribution of Test 2. If $U(x) = 0$ or $U(y) = 0$ the test fails by definition. When we condition on $U(x) = U(y) = 1$, the test distribution is $(A, x, y) \sim \mathcal{D}'$.

Denote by W the set of elements in x_A, by X the set of elements in $x_{\bar{A}}$ and by Y the set of elements in $y_{\bar{A}}$. Let $\pi_1 \in \mathcal{S}_{k/10}$ be the sorting permutation for x_A. Then f, f' satisfy $f'(x)_A = (f(X, W)_W)^{\pi_1}$, and $f'(y)_A = (f(Y, W)_W)^{\pi_1}$. Therefore, $f'(x)_A = f'(y)_A \iff f(X, W)_W = f(Y, W)_W$.

$$\begin{aligned}
\Pr[f' \text{ passes Test 2}] &= \Pr_{A,x,y}[f'(x)_A = f'(y)_A \mid x_A = y_A] \\
&= \Pr_{A,x,y}[U(x) = U(y) = 1 \mid x_A = y_A] \Pr_{(A,x,y)\sim\mathcal{D}'}[f'(x)_A = f'(y)_A] \\
&= (1 - p_1)(1 - p_2) \Pr_{(W,X,Y)\sim\mathcal{D}}[f(X, W)_W = f(Y, W)_W] \\
&= (1 - p_1)(1 - p_2) \Pr[f \text{ passes Item 3 of Test 4}]
\end{aligned}$$

where $\Pr_{A,x,y}[U(x) = U(y) = 1 \mid x_A = y_A] = (1 - p_1)(1 - p_2)$ by the definition of p_1, p_2. □

Claim A.6. *For every function $f : \binom{[N]}{k} \to [M]^k$, the function $f' : [N]^k \to [M]^k \cup \bot$ from Definition A.3 satisfies the following. For every disjoint $W \in \binom{[N]}{k/10}, X \in \binom{[N]}{9k/10}$, every set $A \subset [k], |A| = k/10$ and every pair of permutations $\pi_1 \in S_{k/10}, \pi_2 \in S_{9k/10}$, let $x = ((\vec{W}^{\pi_1})_A, (\vec{X}^{\pi_2})_{\bar{A}})$ where the notation $(y_A, z_{\bar{A}})$ means that we put the elements of y in order in coordinates A, and we put the elements of z in order in the remaining coordinates. Then*

$$\Pr_y[f'(x)_A = f'(y)_A \mid y_A = x_A] = (1 - p_2) \Pr_{Y\sim\mathcal{D}|W,X}[f(X \cup W)_W = f(Y \cup W)_W].$$

Proof. Fix two disjoint subsets $W \in \binom{[N]}{k/10}, X \in \binom{[N]}{9k/10}$, a subset $A \subset [k], |A| = k/10$, and permutations $\pi_1 \in S_{k/10}, \pi_2 \in S_{9k/10}$. Set $x = ((\vec{W}^{\pi_1})_A, (\vec{X}^{\pi_2})_{\bar{A}})$, since X, W are disjoint, $U(x) = 1$.

Let $y \in [N]^k$ be a random string such that $x_A = y_A$. If $U(y) = 0$, then $f'(y) = \bot$ and $f'(x)_A \neq f'(y)_A$. Conditioning on $U(y) = 1$, the distribution over y is $\mathcal{D}'|A, x$. If we take Y to be the elements of $y_{\bar{A}}$, then the distribution over Y is $\mathcal{D}|W, X$.

Clearly, $f'(x)_A = f'(y)_A \iff f(X \cup W)_W = f(Y \cup W)_W$, so

$$\begin{aligned}
\Pr_y[f'(x)_A = f'(y)_A \mid y_A = x_A] &= \Pr_y[U(y) = 0 \mid x_A = y_A] \Pr_{y\sim\mathcal{D}'|A,x}[f'(x)_A = f'(y)_A] \\
&= (1 - p_2) \Pr_{Y\sim\mathcal{D}|W,X}[f(X \cup W)_W = f(Y \cup W)_W].
\end{aligned}$$ □

Proof of Lemma 4.4. Let $f : \binom{[N]}{k} \to [M]^k$ be a function such that $\mathrm{agr}^{Z_{set}}_{k/10}(f) = \epsilon > \mathrm{e}^{-c\alpha k}$, and let $f' : [N]^k \to [M]^k \cup \perp$ be the function from Definition A.3. The constant c is chosen to be small enough so that we can apply Theorem 3.9 with $\alpha = 1/200$. Therefore, we can safely assume that $\alpha \le 1/200$.

By Claim A.5, f' passes Test 2 with probability $\epsilon' = (1-p_1)(1-p_2)\epsilon \ge \mathrm{e}^{-c'\alpha k}$. Theorem 3.9 holds for the function f' with success probability ϵ' and distance parameter α.

By Claim A.6, for every disjoint $W \in \binom{[N]}{k/10}$, $X \in \binom{[N]}{9k/10}$,

$$\Pr_y [f'(x)_A = f'(y)_A \mid y_A = x_A] = (1-p_2) \Pr_{Y \sim \mathcal{D}_{W,X}} [f(X \cup W)_W = f(Y \cup W)_W] .$$

Setting $\eta = 1 - p_1$, this means that if X, W are good, i. e., they satisfy

$$\Pr_Y [f(X \cup W)_W = f(Y \cup W)_W] \ge \eta \frac{\epsilon}{2}$$

then for every set $A \subset [k]$ and permutations π_1, π_2, the restriction $\tau = (A, x_A, f'(x)_A)$ is good, for $x = ((\vec{W}^{\pi_1})_A, (\vec{X}^{\pi_2})_{\bar{A}})$.

Theorem 3.9 states that with probability $1 - \epsilon'^2$ a good restriction is an α-DP restriction. Every pair W, X corresponds to the same number of pairs (A, x), so for at least $(1-\epsilon'^2) \ge (1-\epsilon^2-k^2/N)$ of the pairs W, X, there exists at least one set A and permutations π_1, π_2 such that $\tau = (A, x, f'(x)_A)$ is an α-DP restriction, for $x = ((\vec{W}^{\pi_1})_A, (\vec{X}^{\pi_2})_{\bar{A}})$.

Fix such a pair W, X. We prove that it is an α-DP pair. Let $g_\tau = (g_{\tau,i})_{i \in \bar{A}}$ be the direct product function promised from τ being an α-DP restriction.

We set $g_{W,X} : [N] \to [M]$ to be the following function:

$$\forall a \in [N] \quad g_{W,X}(a) = \text{Plurality}\, \{g_{\tau,i}(a)\}_{i \in \bar{A}} .$$

Let

$$\mathcal{V}_{W,X} = \left\{ Y \in \binom{[N]}{9k/10} \;\middle|\; Y \cap W = \emptyset, f(Y, W)_W = f(X, W)_W \right\}.$$

We show next that $g_{W,X}$ approximates f on $\mathcal{V}_{W,X}$.

Fix $Y \in \mathcal{V}_{W,X}$ such that $g_{W,X}(Y) \overset{3\alpha k}{\not\approx} f(Y \cup W)_Y$. We show that many of the tuples $(x_A, w_{\bar{A}})$ obtained by permuting $Y \cup W$ are "not DP", namely $f(x_A, w_{\bar{A}})_{\bar{A}}$ is far from $g_\tau(w)$ for $w = \vec{Y}^{\pi_3}$ for many choices of the permutation π_3. Since τ is a DP-restriction this must be a rare event, allowing us to upper bound the fraction of such Y.

Let $B \subset Y$ be a set of exactly $3\alpha k$ elements on which $g_{W,X}(Y)$ and $f(Y \cup W)_Y$ differ. That is, $|B| = 3\alpha k$ and for each $b \in B$, $g_{W,X}(b) \ne f(Y \cup W)_b$.

For every $b \in B$, since $g_{W,X}(b)$ is the most frequent value among $g_{\tau,i}(b)$, for at least half of the coordinates $i \in \bar{A}$, $g_{\tau,i}(b) \ne f(Y \cup W)_b$. We call such coordinates i bad (for b), and the rest of the coordinates good. Let π_3 be a random permutation on $9k/10$ elements, so $w = \vec{Y}^{\pi_3}$ places the element b in a random location. Let I_b be the random variable indicating that b is mapped, via π_3, into a good location.

We will show that with high probability, a non-trivial portion of the elements in B are sent to a bad location. For this we will apply the tail bound in Fact 2.6. We prove that $\{I_b\}_{b\in B}$ satisfy the conditions of the fact. Fix a set $S \subsetneq B$, denote $|S| = t$ and denote the elements of S by $b_1 \dots, b_t$ using an arbitrary order. Fix some $b' \in B \setminus S$, then,

$$\begin{aligned}\Pr_{\pi_3}[I_{b'} = 1| \wedge_{i\in[t]} I_{b_i} = 1] &= \sum_{i_1,\dots,i_t\in\bar{A}, i_j\neq i_{j'}} \Pr_{\pi_3}\left[\forall j\in[t], \pi_3(b_j) = i_j\right] \Pr_{\pi_3}\left[I_{b'} = 1|\forall j\in[t], \pi_3(b_j) = i_j\right] \\ &\le \max_{i_1,\dots,i_t\in\bar{A}, i_j\neq i_{j'}} \left\{\Pr_{\pi_3}\left[I_{b'} = 1|\forall j\in[t], \pi_3(b_j) = i_j\right]\right\} \\ &\le \frac{0.5|\bar{A}|}{|\bar{A}| - |S|} \le 0.51.\end{aligned}$$

The last inequality is because $|S| \le |B| \le 3\alpha k$, and we used that $\alpha \le 1/200$. Using this inequality we get that for every $S \subseteq B$, $\Pr_{\pi_3}[\wedge_{b\in S} I_b = 1] \le (0.51)^{|S|}$. By Fact 2.6,

$$\Pr_{\pi_3}\left[\sum_{b\in B} I_b \ge 2\alpha k\right] \le \mathrm{e}^{-3\alpha k\cdot 2\left(\frac{2}{3}-0.51\right)^2} \le \mathrm{e}^{0.14\alpha k}.$$

Whenever π_3 is a permutation with $\sum_{b\in B} I_b < 2\alpha k$, it follows that $f'(x_A, w)_{\bar{A}} \overset{\alpha k}{\not\approx} g_\tau(w)$ (for $w = (\vec{W})^{\pi_3}$). Therefore, we get that

$$\Pr_{Y\in\mathcal{V}_{W,X}}\left[g_{W,X}(Y) \overset{3\alpha k}{\not\approx} f(Y\cup W)_Y\right]\left(1 - \mathrm{e}^{-0.14\alpha k}\right) \le \Pr_{w\in\mathcal{V}_\tau}\left[f'(x_A, w)_{\bar{A}} \overset{\alpha k}{\not\approx} g_\tau(w)\right] \le \epsilon'^2 .$$

This implies that

$$\Pr_{Y\in\mathcal{V}_{W,X}}\left[g_{W,X}(Y) \overset{3\alpha k}{\not\approx} f(Y\cup W)_Y\right] \le \epsilon'^2 + \mathrm{e}^{-0.14\alpha k} \le 2\epsilon^2 . \qquad \square$$

Acknowledgement

The authors wish to thank László Babai for his infinite precision and hyper attention to detail, and for pointing out Bernstein's work that predated Chernoff by decades.

References

[1] SANJEEV ARORA AND MADHU SUDAN: Improved low-degree testing and its applications. *Combinatorica*, 23(3):365–426, 2003. Preliminary version in STOC'97. [doi:10.1007/s00493-003-0025-0, ECCC:TR97-003] 2

[2] VITALY BERGELSON, TERENCE TAO, AND TAMAR ZIEGLER: An inverse theorem for the uniformity seminorms associated with the action of $\mathbb{F}_p^\infty$. *Geom. Funct. Anal. (GAFA)*, 19(6):1539–1596, 2010. [doi:10.1007/s00039-010-0051-1, arXiv:0901.2602] 2

[3] AMEY BHANGALE, IRIT DINUR, AND INBAL LIVNI NAVON: Cube vs. cube low degree test. In *Proc. 8th Innovations in Theoret. Comp. Sci. Conf. (ITCS'17)*, pp. 40:1–31. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. [doi:10.4230/LIPIcs.ITCS.2017.40, arXiv:1612.07491, ECCC:TR16-205] 2

[4] IRIT DINUR: The PCP theorem by gap amplification. *J. ACM*, 54(3):12, 2007. Preliminary version in STOC'06. [doi:10.1145/1236457.1236459, ECCC:TR05-046] 2

[5] IRIT DINUR AND ELAZAR GOLDENBERG: Locally testing direct product in the low error range. In *Proc. 49th FOCS*, pp. 613–622. IEEE Comp. Soc., 2008. [doi:10.1109/FOCS.2008.26] 4, 6

[6] IRIT DINUR, SUBHASH KHOT, GUY KINDLER, DOR MINZER, AND MULI SAFRA: Towards a proof of the 2-to-1 games conjecture? In *Proc. 50th STOC*, pp. 376–389. ACM Press, 2018. [doi:10.1145/3188745.3188804, ECCC:TR16-198] 2

[7] IRIT DINUR AND INBAL LIVNI NAVON: Exponentially small soundness for the direct product Z-test. In *Proc. 32nd Comput. Complexity Conf. (CCC'17)*, pp. 29:1–50. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. [doi:10.4230/LIPIcs.CCC.2017.29] 1

[8] IRIT DINUR AND DAVID STEURER: Direct product testing. In *Proc. 29th IEEE Conf. on Comput. Complexity (CCC'14)*, pp. 188–196. IEEE Comp. Soc., 2014. [doi:10.1109/CCC.2014.27, ECCC:TR13-179] 6, 14, 31, 36

[9] ODED GOLDREICH AND SHMUEL SAFRA: A combinatorial consistency lemma with application to proving the PCP theorem. *SIAM J. Comput.*, 29(4):1132–1154, 2000. Preliminary version in RANDOM'97. [doi:10.1137/S0097539797315744, ECCC:TR96-047] 2

[10] TORBEN HAGERUP AND CHRISTINE RÜB: A guided tour of Chernoff bounds. *Inform. Process. Lett.*, 33(6):305–308, 1990. [doi:10.1016/0020-0190(90)90214-I] 11

[11] WASSILY HOEFFDING: Probability inequalities for sums of bounded random variables. *J. Amer. Statistical Association*, 58(301):13–30, 1963. Also available in The Collected Works of Wassily Hoeffding, 1994, pp. 409–426, Springer and at JSTOR. [doi:10.1080/01621459.1963.10500830] 11, 12

[12] RUSSELL IMPAGLIAZZO AND VALENTINE KABANETS: Constructive proofs of concentration bounds. In *Proc. 14th Internat. Workshop on Randomization and Computation (RANDOM'10)*, pp. 617–631. Springer, 2010. [doi:10.1007/978-3-642-15369-3_46] 10, 12

[13] RUSSELL IMPAGLIAZZO, VALENTINE KABANETS, AND AVI WIGDERSON: New direct-product testers and 2-query PCPs. *SIAM J. Comput.*, 41(6):1722–1768, 2012. Preliminary version in STOC'09. [doi:10.1137/09077299X, ECCC:TR09-090] 2, 5, 6, 7, 8, 13, 28, 30

[14] SUBHASH KHOT, DOR MINZER, AND MULI SAFRA: On independent sets, 2-to-2 games, and Grassmann graphs. In *Proc. 49th STOC*, pp. 576–589. ACM Press, 2017. [doi:10.1145/3055399.3055432, ECCC:TR16-124] 2

[15] MICHAEL MITZENMACHER AND ELI UPFAL: *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge Univ. Press, 2005. [doi:10.1017/CBO9780511813603] 11

[16] ELCHANAN MOSSEL, KRZYSZTOF OLESZKIEWICZ, AND ARNAB SEN: On reverse hypercontractivity. *Geom. Funct. Anal. (GAFA)*, 23(3):1062–1097, 2013. [doi:10.1007/s00039-013-0229-4, arXiv:1108.1210] 8, 12

[17] ALESSANDRO PANCONESI AND ARAVIND SRINIVASAN: Randomized distributed edge coloring via an extension of the chernoff-hoeffding bounds. *SIAM J. Comput.*, 26(2):350–368, 1997. [doi:10.1137/S0097539793250767] 12

[18] RAN RAZ: A parallel repetition theorem. *SIAM J. Comput.*, 27(3):763–803, 1998. Preliminary version in STOC'95. [doi:10.1137/S0097539795280895] 2

[19] RAN RAZ AND SHMUEL SAFRA: A sub-constant error-probability low-degree test, and a sub-constant error-probability PCP characterization of NP. In *Proc. 29th STOC*, pp. 475–484. ACM Press, 1997. [doi:10.1145/258533.258641] 2

AUTHORS

Irit Dinur
Professor
Department of Applied Mathematics
and Computer Science
The Weizmann Institute of Science
Rehovot, Israel
irit.dinur@weizmann.ac.il
http://www.wisdom.weizmann.ac.il/~dinuri/

Inbal Livni Navon
Postdoc
Department of Computer Science
Stanford University
Stanford, CA, USA
inballn@stanford.edu
http://inballn.su.domains/

ABOUT THE AUTHORS

IRIT DINUR is a professor at the Weizmann Institute of Science. She is interested broadly in theoretical computer science and mathematics, and more specifically in complexity theory, probabilistically checkable proofs, hardness of approximation, and most recently in the growing area of high dimensional expansion. She has a wife and three kids.

INBAL LIVNI NAVON is a postdoc at Stanford University. She received her Ph. D. in 2023 from the Weizmann Institute of Science where she was advised by Irit Dinur. She is interested in Algorithmic fairness and in expander graphs and their applications in different areas in theoretical computer science. She is also interested in property testing, error correcting codes and hardness of approximation.