New Direct-Product Testers and 2-Query PCPs

Russell Impagliazzo * Valentine Kabanets † Avi Wigderson ‡

September 30, 2009

Abstract

The "direct product code" of a function f gives its values on all k-tuples $(f(x_1), \ldots, f(x_k))$. This basic construct underlies "hardness amplification" in cryptography, circuit complexity and PCPs. Goldreich and Safra [GS00] pioneered its local *testing* and its PCP application. A recent result by Dinur and Goldenberg [DG08] enabled for the first time testing proximity to this important code in the "list-decoding" regime. In particular, they give a 2-query test which works for *polynomially small* success probability $1/k^\alpha$, and show that no such test works below success probability $1/k$.

Our main result is a 3-query test which works for *exponentially small* success probability $\exp(-k^\alpha)$. Our techniques (based on recent simplified decoding algorithms for the same code [IJKW08]) also allow us to considerably simplify the analysis of the 2-query test of [DG08]. We then show how to *derandomize* their test, achieving a code of polynomial rate, independent of k, and success probability $1/k^\alpha$.

Finally we show the applicability of the new tests to PCPs. Starting with a 2-query PCP over an alphabet Σ and with soundness error $1 - \delta$, Rao [Rao08] (building on Raz's (k-fold) parallel repetition theorem [Raz98] and Holenstein's proof [Hol07]) obtains a new 2-query PCP over the alphabet Σ^k with soundness error $\exp(-\delta^2 k)$. Our techniques yield a 2-query PCP with soundness error $\exp(-\delta\sqrt{k})$. Our PCP construction turns out to be essentially the same as the miss-match proof system defined and analyzed by Feige and Kilian [FK00], but with simpler analysis and exponentially better soundness error.

*U. C. San Diego and Institute for Advanced Study; russell@cs.ucsd.edu

†Simon Fraser University; kabanets@cs.sfu.ca

‡Institute for Advanced Study; avi@ias.edu

Contents

1 Introduction

1.1 Motivation and background

Often in complexity theory, we want to make a somewhat hard problem into a much harder one. One basic tool for doing this is the direct product construction, where the new problem requests answers to a large number (say k) of instances of the original problem. While an intuitive and very useful general method, its correctness (establishing a "direct-product theorem") is frequently non-trivial, often beset with subtleties, and sometimes just wrong. If the answers for the k instances are decided independently, then the solver's probability of success drops exponentially with k. However, sometimes the solver can benefit from using a *correlated* strategy, basing the answer for each instance on the entire set of instances.

A good example is Raz's celebrated parallel repetition theorem [Raz98]. Here, the measure of hardness being improved is the soundness of a probabilistically checkable proof (PCP). Note that the soundness of a PCP often yields a hardness of approximation result for a related problem, so it is very important to get PCPs with optimal soundness. Let us recall how this amplification works. Assume that in the original PCP, on randomness r, the verifier picks two queries at positions x, y of the proof A, and decides according to the "answers" $A[x]$ and $A[y]$. Then the k-fold parallel repetition of that proof system has longer proofs C, indexed by all k-tuples of positions in A, each containing a k-tuple of answers. The new verifier then picks k independent random tapes $r_1, \ldots, r_k$, generating k pairs x_i, y_i, queries the new proof at two positions, obtaining $C[x_1, \ldots, x_k]$ and $C[y_1, \ldots, y_k]$, and finally checks that the original verifier would have accepted for all corresponding pairs of answers to x_i, y_i. Assuming that the acceptance probability of the original verifier was p, how will it drop with this k-fold repetition?

If C was simply A^k, namely if it recorded the answers of A faithfully in all k-tuples, the acceptance probability would drop to p^k. But many counterexamples (see the survey [FV02] and the recent [Raz08]) show that cleverly constructed "proofs" C can in some cases force slower decay in terms of each of the parameters p, k, and moreover must depend on the size of the answer set. These subtleties were so difficult that even showing *any* decay that approaches zero as k increases required a nontrivial proof [Ver96]. A few years later Raz proved his parallel repetition theorem [Raz98], showing that indeed the decay is exponential. Simpler proofs [Hol07, Rao08] and other results give us a pretty good understanding of the limits on the decay in terms of the original parameters, but these remain far from the potentially optimal p^k.

What can be done to salvage the situation and push the soundness amplification towards optimality? (After all, we are the PCP designers, and pure parallel repetitions as above is only one way to go.) Many ideas, both algebraic and combinatorial, were applied to reduce PCP error, and these are beautifully explained in the recent survey of Dinur [Din08]. The best current result is the tour-de-force of Moshkovitz and Raz [MR08]. Here we focus on using direct-product testing for this purpose, an idea pioneered by Goldreich and Safra [GS00]. The idea is to somehow "force" the new proof C to behave like the "direct product" A^k of (some) proof A (or at least reject with high probability those which are not), since if C has this property we could hope for optimal decay.

To compare our and previous results, we view this property as a code. Imagine that (the truth table of) a function $f : \mathcal{U} \to \mathcal{R}$ is encoded by $f^{(k)} : \mathcal{U}^k \to \mathcal{R}^k$, defined by $f^{(k)}(x_1, \cdots x_k) = (f(x_1), \cdots f(x_k))$. Given oracle access to $C : \mathcal{U}^k \to \mathcal{R}^k$, the goal is to test if C is a codeword, or is far from it. In other words, we'd like a test (with few queries to C) that, if passed by a message C with a "significant" probability q, then C is "sufficiently close" to $f^{(k)}$ for some function f. The smaller we can make the value of q that has such implication, the better amplification we can hope for in PCPs.

One should observe immediately that unlike typical error-correcting codes (in particular polynomial-based codes often used in PCPs) this direct-product code is a particularly bad one in standard parameters. For one, its *rate* is lousy – superpolynomial as soon as k is not a constant (we will return to this point when discussing derandomized direct-product codes). For another, its *distance* is even worse – some codewords (e.g., of the Boolean function AND) have exponentially few non-zero entries. Some of the subtleties of direct-product testing arise precisely from these issues. Luckily (and this observation makes the testing possible), for the intended hardness amplification it suffices to certify that, for some f, many entries of C agree with f on many[1] (rather than all) of the k answers. In other words, C must be close to an *approximate* direct-product codeword. With that notion of "proximity" or "decoding" in mind, one tries to devise a test to certify it for small success probability q, hopefully approaching the optimal p^k. We note that such "proximity testers" were formalized in a general setting under the name "spot-checkers" in [EKK$^+$00].

Initial work addressed the case in which the success probability q of the test is very close to 1. This is sometimes called the "unique decoding" regime, since in this case it is possible to show that "decoded" function f is unique. The original paper [GS00] described a test with a constant number of queries, and this was improved to the optimal two-query test by Dinur and Reingold [DR06]. Even for these results, with q extremely high, the proofs are quite nontrivial.

But for PCPs with small soundness error we need to tackle small q, and one can easily see that as soon as $q \leqslant 1/2$ unique decoding is impossible. Indeed, let C agree with each of t direct product codewords $f_i^{(k)}$ in a q-fraction of its coordinates, for some (random) functions $f_1, \ldots, f_t$ and t about $1/q$. Thus if C passes the test with a small probability q, the best "explanation" we can hope for is such a short list of codewords, i.e., a "list-decoding algorithm" rather than unique decoding. In general, list-decoding of codes has been very important in recent developments in coding theory and complexity theory, but seems to require more subtlety in algorithm design and analysis than unique decoding.

The first result to test the direct-product code in the list-decoding regime was obtained by Dinur and Goldenberg [DG08] (building on the earlier work by Feige and Kilian [FK00]). They give a 2-query test which, if C passes with probability $q > 1/k^\alpha$ (for some fixed $\alpha > 0$), certifies that C is close to some codeword. The proof is quite involved. Moreover, they dash the hope of achieving exponential decay of q in terms of k, showing it impossible for 2-query tests even for q that is inverse polynomial in $1/k$.

1.2 Our results

Our main result is that only one additional query is needed to go from polynomially to exponentially small error. We give a 3-query test which, if passed by C with probability $q > \exp(-k^{1/3})$, certifies that C approximately agrees with a direct product function on a $poly(q)$-fraction of its entries. Our techniques (see below) also allow us to considerably improve the analysis of the 2-query test of [DG08] (for poly$(1/k)$-agreement).

To explain our next result, *derandomized* direct product testing, we revisit the PCP motivation. Another important parameter for applications to PCPs is the *proof size*. In coding terms, the proof size is inversely proportional to the *rate* of the code. Note that the k-fold direct-product code blows up the "message" (namely the truth table of f, which would be the original PCP size) to the kth power. To achieve subconstant soundness q, even assuming optimal decay $q = p^k$, we must take k to be nonconstant, which immediately makes the proof size superpolynomial[2]. A natural way

[1] In the PCP application, "many" means a p-fraction, where p is the success probability of the original verifier

[2] So using this construction, we cannot get inapproximability results based only on the assumption $\mathsf{P} \neq \mathsf{NP}$, but

around this is to have the encoding of f provide its values not on *all* k-tuples, but rather on a much smaller subset of these tuples. The hope would be that such small (but carefully chosen) subset will still allow testing, and hence PCPs with improved soundness.

Goldreich and Safra [GS00] gave the first derandomized direct product test in the unique decoding regime (for constant acceptance probability ϵ), using a constant number of queries. The possibility of a derandomized *2-query* test (even in the unique decoding regime) was raised in [DG08] as an open question. We not only solve this question in the unique decoding regime, but also in the list-decoding regime. We show that for any k, there is a family of k-tuples of size a fixed polynomial in n which is independent of k, so that if C passes the 2-query test of [DG08] with probability $q > 1/k^{\alpha}$ then it must have poly(q)-agreement with an approximate direct-product codeword. In coding language, we provide a locally testable, approximate list-decodable k-fold direct-product code of inverse polynomial rate.

Finally, we return to the motivation of using direct-product testing to improve the soundness amplification of PCPs. In PCPs, there is a big gap between two and more than two queries, in terms of the naturalness of the consequent constraint satisfaction problems one gets hardness of approximation for. If we combined a 3-query direct-product test with a 2-query parallel repetition, that would seem to suggest we would only get a 5-query PCP of dubious value. Moreover, thinking through the requirements of PCP proofs more closely, they do not seem to match those of DP-testing. In the list decoding regime, closeness to some codeword is not actually the property that we want for PCPs. The existence of *one* codeword that agrees with our message a non-negligible fraction of the time doesn't guarantee that almost all of the rest of the time, the prover isn't getting the advantage of a correlated strategy. (This is not an issue in the unique decoding world, since there the proof must be close to a single direct product function almost everywhere.) We need that, *conditional* on the proof passing our test, *almost surely* the proof is close to a direct product. In this sense, our original goal of testing was too modest.

On the other hand, it is actually not important that there be a single direct product function, as opposed to a distribution of such functions (independent of the query), maybe even one where no element appears very often. Since soundness improves exponentially for each direct product in the support of the distribution, it improves similarly for the entire distribution. In this sense, our testing condition is too strong for the original application.

Fortunately, while the *existence* of a 3-query DP-test doesn't seem helpful for PCPs, our *analysis* of the 3-query test is applicable. In particular, we show that, even when the 2-query test is useless as a direct product tester, it is useful to certify that a function C is close to a *distribution of direct products*. Our 3-query test then follows as a consequence, as does the use of the 2-query test when q is polynomially large. However, this kind of *distributional direct product testing* is actually what we need for PCPs (and is easy to merge with the parallel repetition of the proof without additional queries). We show, as a "proof of concept", a general construction improving the soundness of a PCP from $1 - \delta$ to $\exp(-\delta\sqrt{k})$ that makes only two queries. Our PCP construction turns out to be closely related to the 2-prover protocol defined and analyzed by Feige and Kilian [FK00]. Our analysis, however, yields a much better (exponential, as opposed to polynomial) decay in the number k of repetitions, and is arguably simpler than that of [FK00]. With current technology, the construction of [Rao08], using an improved analysis of parallel repetition theorem [Raz98, Hol07] for a subclass of games, is superior to ours. However, we see no reason in principle why our test should not be improvable to have better decay, or even a derandomized variant. Clarifying the limits of our approach, compared with parallel repetition, is an extremely interesting direction.

must make stronger assumptions, such as NP $\not\subseteq$ quasi-P.

1.3 Our techniques, and direct-product decoding

The direct-product construction has long been central in complexity theory and cryptography. Yao's XOR Lemma [Yao82, Lev87], and its sibling, the "concatenation lemma" (provably equivalent using the results of Goldreich and Levin [GL89]), are the basic hardness amplification tools in these areas. These two theorems have many many different proofs (e.g,. [GNW95, IW97]), with different parameters, and which each have found different extensions and generalizations.

Impagliazzo [Imp02] and Trevisan [Tre03] reformulated the combinatorial heart of the concatenation lemma in the language of coding theory, as an "approximate list-decoding" problem: Given a corrupted direct product C, with the promise that it has q-agreement with some direct-product function f^k, find a list of functions $f_1, ...f_l$ so that any possible f is close to at least one of the f_i's (in Hamming distance). Trevisan [Tre03] observed that the list-size of such an algorithm quantifies the non-uniformity of the proof, and used this connection for hardness amplification versus uniform adversaries. [IJK06, IJKW08] improved the list-size over previous proofs, to almost the information-theoretic optimal value.

There is no clear reduction between direct product testing and direct product decoding.[3] In direct product decoding, you are guaranteed that a function is close to a direct product; in testing, you wish to decide whether this is the case. In decoding, you need to find the function; in testing, you simply need to accept or reject. Finally, in decoding, you typically are allowed a number of queries that is polynomial in the agreement parameter. In testing, it is vital to absolutely minimize the number of queries, ideally with a small number that does not depend on the agreement at all. Despite these differences, there seem to be deep connections between the two concepts. In particular, testing almost always seems harder, with an empirical reason being that essentially the only way to analyze a test is to show how it decodes a small list.

In the past couple of years we have been developing (with Jaiswal) [IJK06, IJKW08] a set of tools which allowed us to get optimal list-decoding of the direct product code, as well as to derandomize some of its versions (for the purpose of decoding). A central part of that work, as is of all mentioned work on testing, is understanding the following, extremely natural 2-query test applied to an oracle C: Pick two k-tuples at random, under the condition that they agree on some subset of size k' of the coordinates. The main question is what structural information can be obtained about C if it passes the test (namely answers consistently on the common queries) with probability q. Precisely such structural information is obtained in the decoding papers. This current work draws much from these, and adapts them to the testing problem. As explained above, in the testing world one wants to certify what is given as an assumption in the decoding world, and so this adaptation is sometimes impossible (as the [DG08] counterexample shows) and sometimes possible but intricate. But many of the technical notions and lemmas nevertheless apply here. We feel that clarifying the connections between the testing and decoding problems will be extremely enlightening.

1.4 Formal statements of our main results

DP testing. Here we formally state our direct product testing results. Let C be a given oracle (circuit) that presumably computes the direct product f^k, for some function $f : \mathcal{U} \to \mathcal{R}$.[4] It will be more convenient for us to view the k-wise direct product as defined over *sets* of size k, rather than ordered k-tuples; however, our results work for k-tuples as well.

[3] Our comments below also apply to testing/decoding of other codes (and properties).

[4] Think of Boolean functions f for simplicity. However, Section 5 shows our tests work for arbitrary ranges $\mathcal{R}$.

We will argue that the following 3-query test, which we call a Z-test, can certify this. Below, for disjoint sets A and B, we denote by (A, B) the union $A \cup B$. Also, for $A \subset S$, we denote by $C(S)|_A$ the answers $C(S)$ for the subset A.

Z-Test:
1. Pick a random k-set $(A_0, B_0) \subseteq \mathcal{U}$, where $|A_0| = k' = \Theta(\sqrt{k})$.
2. Pick a random set $B_1 \subseteq \mathcal{U} \setminus A_0$ of size $k - k'$. If $C(A_0, B_0)|_{A_0} \neq C(A_0, B_1)|_{A_0}$, then reject; otherwise continue.
3. Pick a random set $A_1 \subseteq \mathcal{U} \setminus B_1$ of size k'. If $C(A_0, B_1)|_{B_1} \neq C(A_1, B_1)|_{B_1}$, then reject; otherwise, accept.

The test above makes 3 queries to the oracle C, and makes two checks for agreement: first on a subset A_0, then on a subset B_1. If we restrict this test to just the first two steps, we get the following 2-query test analyzed by [DR06, DG08].

V-Test:
1. Pick a random k-set $(A_0, B_0) \subseteq \mathcal{U}$, where $|A_0| = k' = \Theta(\sqrt{k})$.
2. Pick a random set $B_1 \subseteq \mathcal{U} \setminus A_0$ of size $k - k'$. If $C(A_0, B_0)|_{A_0} \neq C(A_0, B_1)|_{A_0}$, then reject; otherwise accept.

Intersecting sets chosen in the 3-query Z-test and the 2-query V-test can be pictured to form the letters "Z" and "V", respectively, see Fig. 1 below, – whence the names of these tests.

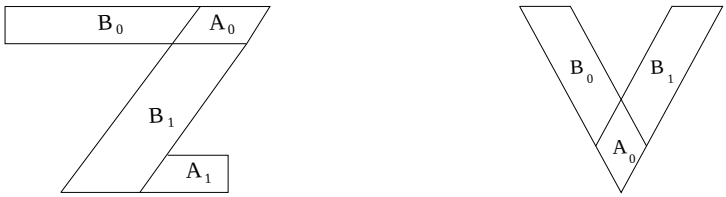

Figure 1: Z-test and V-test pictorially.

As proved by [DG08], the V-test is useless for the acceptance probability below $1/k$. Here we show that, with just one extra query, the resulting Z-test is useful even for inverse-exponentially small acceptance probability. For the proof of the following theorem, see Section 3.

Theorem 1.1 (DP Testing)**.** *There are constants $0 < \eta_1, \eta_2 < 1$ such that, if the Z-test accepts with probability ϵ, for $\epsilon > e^{-k^{\eta_1}}$, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that, for each of at least $\epsilon/4$ fraction of k-sets S from $\mathcal{U}$, the oracle value $C(S)$ agrees with the direct product $g^k(S)$ for all but at most $k^{-\eta_2}$ fraction of elements in S.*

Next we describe our derandomized DP test. We define the derandomized direct product similarly to [IJKW08]. Let $k = q^d$ for some prime power q, and some constant d (to be determined). We identify the domain $\mathcal{U}$ with some m-dimensional linear space over the field $\mathbb{F}_q$, i.e., $\mathcal{U} = \mathbb{F}_q^m$. The k-wise direct product of a function $f : \mathcal{U} \to \mathcal{R}$ is defined as follows: Given a d-dimensional linear[5] subspace A of $\mathcal{U}$, we set $f^k(A)$ to be the values of f on all $k = q^d$ points in the subspace A (ordered according to some fixed ordering of $\mathcal{U}$). For subspaces A and B of $\mathcal{U}$, we denote by $A + B$ the set $\{a + b \mid a \in A, b \in B\}$, where $a + b$ means component-wise addition of the vectors a and b.

The following is an analogue of the Z-test for the derandomized case.

[5] [IJKW08] uses *affine* subspaces, but one could also use just linear subspaces, with a tiny loss in parameters.

Derandomized Z-Test:
1. For $d_0 = d/25$ (for some constant $d \geqslant 25$), pick a random d_0-dimensional subspace A_0, and a random $(d-d_0)$-dimensional subspace B_0 of $\mathcal{U}$ that is linearly independent from A_0.
2. Pick a random $(d-d_0)$-dimensional linear subspace B_1 of $\mathcal{U}$ that is linearly independent from A_0. If $C(A_0+B_0)|_{A_0} \neq C(A_0+B_1)|_{A_0}$, then reject; otherwise, continue.
3. Pick a random d_0-dimensional subspace A_1 linearly independent from B_1. If $C(A_0+B_1)|_{B_1} \neq C(A_1+B_1)|_{B_1}$, then reject; otherwise, accept.

We prove the following (see Section 4 for the proof).

Theorem 1.2 (Derandomized DP Testing)**.** *There are constants $0 < \eta_1, \eta_2 < 1$ such that, if the derandomized Z-test accepts with probability ϵ, for $\epsilon \geqslant k^{-\eta_1}$, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that, for each of at least $\epsilon/4$ fraction of d-dimensional subspaces S from $\mathcal{U}$, the oracle value $C(S)$ agrees with the direct product $g^k(S)$ for all but at most $k^{-\eta_2}$ fraction of elements in S.*

Our techniques also allow us to get a simpler analysis of the V-test for the case of acceptance probability $\epsilon > \text{poly}(1/k)$, first shown by [DG08]; see Section 3.4 for the proof. Moreover, the same analysis shows that the *derandomized* V-test (the first two steps of the derandomized Z-test) also works; see Section 4 for the proof.

Theorem 1.3. *There is a constant $0 < \eta < 1$ such that, if the derandomized V-test accepts with probability $\epsilon \gg \sqrt{k'/k}$, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that for at least $\epsilon' = \Omega(\epsilon^6)$ fraction of subspaces S, the oracle $C(S)$ agrees with $g(S)$ in all but at most $k^{-\eta}$ fraction of inputs $x \in S$.*

We remark that, in both independent and derandomized cases, we also get approximate, local, list-decoding algorithms for the corresponding DP codes.

PCP. As another application of our techniques, we get a generic reduction from 2-query PCPs, over an alphabet Σ with completeness σ and soundness $1-\delta$, to 2-query PCPs, over the alphabet Σ^k with completeness $1-\exp(-\sigma k)$ and soundness $\exp(-\delta k')$, for $k' = \Theta(\sqrt{k})$. Our reduction preserves *perfect* completeness: if the initial PCP has $\sigma = 1$, then so does the resulting PCP. We describe this construction next.

Consider a constraint satisfaction problem (CSP) for regular undirected graphs, over an alphabet Σ. An instance of such a CSP consists of a regular undirected graph $G = (U, E)$ on n nodes and a family $\Phi = \{\phi_e\}_{e \in E}$ of constraints, where each edge $e = (x, y) \in E$ has an associated constraint $\phi_e : \Sigma^2 \to \{0, 1\}$ (which need not be symmetric). For $0 \leqslant \sigma, \delta \leqslant 1$, a CSP instance is *$\sigma$-satisfiable* if there is an assignment $f : U \to \Sigma$ that satisfies at least σ fraction of edge constraints; a CSP instance is *δ-unsatisfiable* if every assignment $f : U \to \Sigma$ violates at least δ fraction of edge constraints.

Given a CSP-instance (G, Φ) (where G is a regular undirected graph on n nodes), we will ask for an assignment C_E that, given a set of k edges in the constraint graph G, returns assignments to all of the end-points of these edges. We give a 2-query verifier that almost certainly accepts an honest proof C_E for a σ-satisfiable CSP instance, and almost certainly rejects any proof for a δ-unsatisfiable CSP instance, where the rejection probability is independent of the size of the alphabet Σ.

Let $k' < k$ be the parameter from our DP test above. Our 2-query verifier is the following.

Verifier $\mathcal{V}$:
1. Pick a set of k' random vertices A. For each vertex $v \in A$, pick a random incident

edge (v, v') in G. Let $A_{E,1}$ be the set of these k' edges. Independently, pick another set $A_{E,2}$ of k' random edges incident on the vertices in A. Finally, pick two random sets of edges $B_{E,1}$ and $B_{E,2}$, of size $k - k'$ each.
2. Query $C_E(A_{E,1}, B_{E,1})$ and $C_E(A_{E,2}, B_{E,2})$. Accept iff the following checks pass:
(a) the query answers satisfy $0.9 \cdot \sigma$ fraction of constraints on each of the $B_{E,i}$'s[6], and
(b) they assign the same values to A.

Theorem 1.4. *(i) If a CSP-instance (G, Φ) is σ-satisfiable, then there is a proof C_E accepted by verifier $\mathcal{Y}$ with probability $\sigma' \geqslant 1 - \exp(-\sigma k)$; moreover, if $\sigma = 1$, then $\sigma' = 1$. (ii) There is a constant $c > 0$ such that, if the CSP-instance is δ-unsatisfiable, then no proof C_E is accepted by $\mathcal{Y}$ with probability greater than $\epsilon = e^{-(1/c)\delta k'}$, provided that $\epsilon < 1/4$.*

The proof of this theorem is given in Section 6. Together with the PCP Theorem [AS98, ALM+98] (e.g., using [Din07]), but *without* the parallel repetition theorem of [Raz98], Theorem 1.4 implies that NP has 2-query PCPs with perfect completeness, soundness $\exp(-\sqrt{k})$, and proof size $n^{O(k)}$. In fact, this theorem can be interpreted as a new parallel repetition theorem for certain 2-prover games, where the value of the repeated game decreases exponentially with the number of repetitions, independent of the alphabet size; see Theorem 6.3 in Section 6.

Before Raz's celebrated result [Raz98], Feige and Kilian [FK00] and Verbitsky [Ver96] gave the first proofs that (some version of) parallel repetition indeed decreases the soundness of 2-prover games. It turns out that our techniques yield a significantly improved analysis of the construction from [FK00]. More precisely, we can analyze the following 2-prover protocol, which is essentially the same as the *miss-match* proof system introduced by Feige and Kilian [FK00].

As before, let (G, Φ) be a regular graph CSP with the vertex set U and the alphabet Σ. The first prover C_1 gets as input a k'-subset of vertices of G and returns an assignment to all these vertices. The second prover is a function C_E that, given a set of k edges of G, returns assignments to all the $2k$ end-points of these edges. Consider the following protocol.

Verifier $\mathcal{Y}'$:
1. Pick a set of k' random vertices A. For each vertex $v \in A$, pick a random incident edge (v, v') in G. Let $A_{E,2}$ be the set of these k' edges. Pick a set of $(k - k')$ random edges $B_{E,2}$.
2. Query $C_1(A)$ and $C_E(A_{E,2}, B_{E,2})$. Accept iff the following checks pass:
(a) the query answers satisfy $0.9 \cdot \sigma$ fraction of constraints of $B_{E,2}$, and
(b) they assign the same values to A.

The advantage of $\mathcal{Y}'$ over $\mathcal{Y}$ is that $\mathcal{Y}'$ satisfies the *projection property*: the answers of the prover C_E determine the answers of the prover C_1. We prove in Section 6.3 that $\mathcal{Y}'$ has soundness $\exp(-\delta k')$; in contrast, the analysis of [FK00] yields only inverse polynomial soundness.

Theorem 1.5. *(i) If a CSP-instance (G, Φ) is σ-satisfiable, then there are proofs (C_1, C_E) accepted by verifier $\mathcal{Y}'$ with probability $\sigma' \geqslant 1 - \exp(-\sigma k)$; moreover, if $\sigma = 1$, then $\sigma' = 1$. (ii) There is a constant $c > 0$ such that, if the CSP-instance is δ-unsatisfiable, then no proofs (C_1, C_E) are accepted by $\mathcal{Y}'$ with probability greater than $\epsilon = e^{-(1/c)\delta k'}$, provided that $\epsilon < 1/4$.*

We see no reason why the exponential decay of the PCP constructions above cannot be improved to $\exp(-\delta k)$.

[6] Actually, we only need this for $B_{E,2}$.

Remainder of the paper. We give the definitions of inclusion graphs and prove their sampling properties in Section 2. We prove Theorem 1.1 in Section 3, and Theorem 1.2 in Section 4. In Section 5 we prove that all our direct product testing results hold for functions with arbitrary (not necessarily Boolean) range. We prove Theorems 1.4 and 1.5 in Section 6.

2 Preliminaries

2.1 Notation

For a natural number $n \in \mathbb{N}$, we denote by $[n]$ the set $\{1, 2, \ldots, n\}$. For $0 \leqslant \alpha \leqslant 1$, and k-tuples a and b, we write $a \overset{>\alpha}{\neq} b$ to denote that a and b differ in more than α fraction of positions.

For a graph G and a vertex v of G, we denote by $N_G(v)$ the set of all neighbors of v in G; usually we will drop the subscript G if the graph is clear from the context.

2.2 Linear spaces

We will need the following simple lemma.

Lemma 2.1. *For any integers $c, d > 0$ and $D = cd$, and a field $\mathbb{F}_q$, the D-dimensional linear space $\mathbb{F}_q^D$ has $t = (q^D - 1)/(q^d - 1)$ linear d-dimensional subspaces that are pairwise disjoint except for the common zero vector.*

Proof. Let $Q = q^d$. The D-dimensional vector space $\mathbb{F}_q^D = \mathbb{F}_q^{dc}$ can be viewed as $\mathbb{F}_{q^d}^c$, the c-dimensional space over the field $\mathbb{F}_Q$. The c-dimensional space $\mathbb{F}_Q^c$ has exactly $(Q^c - 1)/(Q - 1) = t$ distinct lines through 0 (i.e., 1-dimensional linear subspaces). Consider any such line in $\mathbb{F}_Q^c$. The points on the line are given by some equation of the form $\vec{a} \cdot x$, where $\vec{a} \in \mathbb{F}_Q^c$ is some non-zero vector, and x is a variable assuming values in $\mathbb{F}_Q$. Each point on the line is an element of $\mathbb{F}_Q^c$, and so corresponds to a vector in $\mathbb{F}_q^D$. Using the correspondence between $\mathbb{F}_Q$ and $\mathbb{F}_q^d$, it is easy to show that the collection of points on the given line $\vec{a} \cdot x$ corresponds to a linear subspace of $\mathbb{F}_q^D$. Since the line has exactly $Q = q^d$ distinct point, we get that the dimension of this linear subspace over $\mathbb{F}_q$ is d. Since any two distinct lines through 0 share only the zero vector, the lemma follows. □

We will also need the following sampling property of random linear subspaces (based on pairwise independence); this result is implicit in the journal version of [IJKW08] (to appear in SIAM Journal on Computing).

Lemma 2.2. *Let V_0 and V_1 be a pair of arbitrary linear spaces over a field $\mathbb{F}_q$ that are disjoint except for the common zero vector. Let $X \subseteq V_0 + V_1$ be any subset of points of measure μ. Finally, let $W \subset V_1$ be a random linear subspace of V_1. Then*

$$\mathbf{Pr}\left[\left|\frac{|(V_0 + W) \cap X|}{|V_0 + W|} - \mu\right| \geqslant \mu/2\right] \leqslant \frac{4q^2}{|W|\mu}.$$

2.3 Sampler graphs

For our analysis of DP tests, we use basic sampling lemmas, which (as in [IJKW08]) can be stated in the graph-theoretic language. Let $G(L, R) = (L \cup R, E)$ be a bipartite bi-regular graph, where we think of L as left vertices, and R as right vertices. For $0 \leqslant \alpha, \beta \leqslant 1$, we call G a (α, β)-*sampler* if, for every subset $F \subseteq L$ of measure $\mu \geqslant \alpha$, there are at most $\beta|R|$ vertices $r \in R$ where

$\left|\mathbf{Pr}_{\ell \in N(r)}[\ell \in F] - \mu\right| \geqslant \mu/2$. *Inclusion graphs* are graphs whose vertices are subsets of some finite universe, and two vertices (subsets) are connected by an edge iff one is contained in the other. We usually think of these inclusion graphs as bi-partite, with smaller subsets as the left vertices.

Let $\mathcal{U}$ be a finite universe. We will need the following inclusion graphs $G(L, R)$:

- **Independent:** L are all s_l-subsets of $\mathcal{U}$, and R are all s_r-subsets of $\mathcal{U}$, where $s_r = t \cdot s_l$ for an integer $t > 1$.
- **Subspaces:** For $\mathcal{U} = \mathbb{F}_q^m$, L are all d_l-dimensional linear subspaces of $\mathcal{U}$, and R are all d_r-dimensional linear subspaces of $\mathcal{U}$, where $d_r = c \cdot d_l$ for an integer $c > 1$.

We show that these inclusion graphs are samplers.

Lemma 2.3 (Subset/Subspace Samplers)**.** *Both Independent and Subspaces inclusion graphs $G(L, R)$ defined above are (α, β)-samplers, where*

- ***Independent:*** *$\beta = e^{-\Omega(\alpha t)}$, provided that $(t \cdot s_l)^2/|\mathcal{U}| \leqslant e^{-\Omega(\alpha t)}$ and $\alpha t \geqslant \Omega(\ln 1/\alpha)$.*
- ***Subspaces:*** *$\beta = O(1/\sqrt{\alpha q^{(c-1)d_l}})$, provided that $\alpha^{3/2} q^{(c-1)d_l/2} > 10$.*

Proof. **Independent:** Let $M = s_r$ be the size of subsets on the right side of the bipartition of $G(L, R)$. Let $S_1, \ldots, S_t$ be any fixed partition of the set $[M]$ into t subsets of size s_l each; e.g., $S_i = \{s_l(i-1)+1, \ldots, s_l i\}$ for $1 \leqslant i \leqslant t$. Let $\mathbf{G} = \mathbb{S}_M$ be the permutation group on $[M]$.

Every M-subset B of $\mathcal{U}$ can be viewed as an ordered M-tuple, according to some fixed (say, lexicographical) ordering of the universe $\mathcal{U}$. Thus we can index the elements in B by elements of $[M]$. For every subset $S \subseteq [M]$, let $S(B)$ denote the subset of the elements of B whose indices are in the set S.

Observe that for every fixed S_i, a random permutation $\pi \in \mathbf{G}$ maps S_i to a uniformly random s_l-subset πS_i of $[M]$. Hence, for every fixed S_i, if we pick a random M-subset B of $\mathcal{U}$ and a random permutation $\pi \in \mathbf{G}$, we get that $(\pi S_i)(B)$ is a uniformly random s_l-subset of $\mathcal{U}$.

Let $L' \subseteq L$ be any subset of measure $\lambda \geqslant \alpha$. By the above, we get that

$$\lambda = \mathbf{Pr}_{i \in [t], B \in R, \pi \in \mathbf{G}}[(\pi S_i)(B) \in L'] = \mathbf{Exp}_{B \in R, \pi \in \mathbf{G}}\left[\mathbf{Pr}_{i \in [t]}[(\pi S_i)(B) \in L']\right].$$

For a random permutation $\pi \in \mathbf{G}$ and a random $B \in R$, the subsets $(\pi S_1)(B), \ldots, (\pi S_t)(B)$ are distributed as a uniform t-tuple of pairwise disjoint s_l-subsets of $\mathcal{U}$. This is essentially the same as a uniformly chosen t-tuple of elements of L. Indeed, suppose we pick $S'_1, \ldots, S'_t$ uniformly from L. The probability of some pair S'_i and S'_j having a nonempty intersection is at most t^2 times the probability that two random s_l-size subsets of $\mathcal{U}$ have nonempty intersection. By Markov's inequality, the latter is at most $s_l^2/|\mathcal{U}|$, and so the overall statistical distance between the two distributions on $S'_1, \ldots, S'_t$ is at most $(ts_l)^2/|\mathcal{U}|$, which is at most $e^{-\Omega(\alpha t)}$ by our assumption.

Hence, we have

$$\mathbf{Pr}_{\pi \in \mathbf{G}, B \in R}[|\mathbf{Pr}_{i \in [t]}[(\pi S_i)(B) \in L'] - \lambda| \geqslant \lambda/3] \leqslant \mathbf{Pr}_{S'_1, \ldots, S'_t \subset L}[|\mathbf{Pr}_{i \in [t]}[S'_i \in L'] - \lambda| \geqslant \lambda/3] + (ts_l)^2/|\mathcal{U}|.$$

By the Chernoff bound, $\mathbf{Pr}_{S'_1, \ldots, S'_t \subset L}[|\mathbf{Pr}_{i \in [t]}[S'_i \in L'] - \lambda| \geqslant \lambda/3] \leqslant e^{-\Omega(\lambda t)} \leqslant e^{-\Omega(\alpha t)}$. So, for $p = e^{-\Omega(\alpha t)} + (ts_l)^2/|\mathcal{U}| \leqslant e^{-\Omega(\alpha t)}$, we get that $\mathbf{Pr}_{\pi \in \mathbf{G}, B \in R}[|\mathbf{Pr}_{i \in [t]}[(\pi S_i)(B) \in L'] - \lambda| \geqslant \lambda/3] \leqslant p$. By averaging, we get that for at least $1 - \sqrt{p}$ of the sets $B \in R$, it is the case that for at least $1 - \sqrt{p}$ of $\pi \in \mathbf{G}$, the fraction of subsets $(\pi S_i)(B)$ that fall into L' is between $(2/3)\lambda$ and $(4/3)\lambda$.

Finally, for a given $B \in R$, the probability that a random s_l-subset of B falls into L' is $\mathbf{Exp}_{\pi \in \mathbf{G}}\left[\mathbf{Pr}_{i \in [t]}[(\pi S_i)(B) \in L']\right]$. By the above, for all but at most $\sqrt{p}$ fraction of sets B, this

average over $\pi \in \mathbf{G}$ will be at least $(1-\sqrt{p})(2/3)\lambda \geqslant \lambda/2$ and at most $(4/3)\lambda + \sqrt{p} \leqslant (3/2)\lambda$, since $\sqrt{p} \leqslant \lambda/10$ (by our assumption that $\alpha t \geqslant \Omega(\ln 1/\alpha)$).

Subspaces: The proof is similar to that for the case of Independent, except we'll be using pairwise independence and the Chebyshev bound (rather than full independence and the Chernoff-Hoeffding bound). Let $D = d_r$. For $t = (q^D - 1)/(q^{d_l} - 1)$, let $S_1, \ldots, S_t$ be any fixed collection of d_l-dimensional linear subspaces of $\mathbb{F}_q^D$ that are pairwise disjoint except for the common zero, as guaranteed by Lemma 2.1. Let $\mathbf{G} = \mathbb{GL}(D, q)$, i.e., the matrix group of all nonsingular $D \times D$ matrices over $\mathbb{F}_q$.

A random D-dimensional subspace B of $\mathcal{U}$ is specified by a random set of D linearly independent (basis) vectors from $\mathcal{U}$. For any d_l-dimensional subspace S of $\mathbb{F}_q^D$, let $S(B)$ denote the corresponding d_l-dimensional subspace in B. Clearly, the subspaces $S_i(B)$ and $S_j(B)$ have only the zero vector in common, for any $1 \leqslant i \neq j \leqslant t$.

Observe that for each fixed d-dimensional linear subspace S of $\mathbb{F}_q^D$, applying a random linear transformation $A \in \mathbf{G}$ to S results in a uniformly distributed d-dimensional linear subspace AS of $\mathbb{F}_q^D$. Hence, for every fixed $1 \leqslant i \leqslant t$, the subspace $(AS_i)(B)$ is uniform over L, for randomly chosen $B \in R$ and $A \in \mathbf{G}$. Moreover, for every pair of indices $1 \leqslant i \neq j \leqslant t$, if we pick random $B \in R$ and $A \in \mathbf{G}$, we get that the linear subspaces $(AS_i)(B)$ and $(AS_j)(B)$ are nearly independent in the following sense: the pair $((AS_i)(B), (AS_j)(B))$ is uniform over all pairs of *linearly independent* d_l-dimensional subspaces in L. Since the probability of picking two linearly dependent d_l-dimensional subspaces in $\mathcal{U}$ is at most q^{2d_l}/q^m, which is negligible, we will essentially be able to assume that the sequence $(AS_1)(B), \ldots, (AS_t)(B)$ is a sequence of pairwise independent random elements in L.

Let $L' \subseteq L$ be any subset of measure $\lambda \geqslant \alpha$. The probability that a random d_l-dimensional subspace of a random D-dimensional subspace $B \in R$ falls into L' is

$$\lambda = \mathbf{Pr}_{B \in R, A \in \mathbf{G}, i \in [t]}[(AS_i)(B) \in L'] = \mathbf{Exp}_{A \in \mathbf{G}, B \in R}\left[\mathbf{Pr}_{i \in [t]}[(AS_i)(B) \in L']\right].$$

We will use Chebyshev's inequality:

$$\mathbf{Pr}_{B \in R, A \in \mathbf{G}}\left[\left|\mathbf{Pr}_{i \in [t]}[(AS_i)(B) \in L'] - \lambda\right| \geqslant \lambda/3\right] \leqslant \mathbf{Var}_{B \in R, A \in \mathbf{G}}\left[\sum_{i=1}^{t} \chi[(AS_i)(B) \in L']\right] / (t^2\lambda^2/9), \tag{1}$$

where χ is the indicator function such that $\chi[E]$ is 1 if an event E occurs, and is 0 otherwise.

To simplify the notation, let us denote by X_i the random variable (of B and A) that is 1 if the subspace $(AS_i)(B) \in L'$, and 0 otherwise. Let $X = \sum_{i=1}^{t} X_i$. We have that $\mathbf{Exp}[X] = t\lambda$, and $\mathbf{Var}[X] = \mathbf{Exp}[X^2] - (t\lambda)^2$. The expectation of X^2 is $\mathbf{Exp}[X^2] = t\lambda + 2\sum_{i<j} \mathbf{Exp}[X_i \cdot X_j]$.

To bound the probability $\mathbf{Pr}[X_i = 1 \ \wedge \ X_j = 1] = \mathbf{Pr}[X_i = 1] \cdot \mathbf{Pr}[X_j = 1 | X_i = 1]$, for each $i < j$, we use the "near" pairwise independence of the subspaces $(AS_i)(B)$ and $(AS_j)(B)$ mentioned above. Namely, for each linear subspace $S \in L'$, we have that, conditioned on random $B \in R$ and $A \in \mathbf{G}$ such that $(AS_i)(B) = S$, the subspace $(AS_j)(B)$ is uniform over all d_l-dimensional subspaces of $\mathcal{U}$ that are disjoint from S (except for the common zero vector). Since all but at most $\tau = q^{2d_l}/q^m$ fraction of d_l-dimensional subspaces of $\mathcal{U}$ are disjoint from S, we get that the conditional probability distribution of $(AS_j)(B)$ (for random $B \in R$ and $A \in \mathbf{G}$ such that $(AS_i)(B) = S$) is at most the statistical distance τ away from the uniform distribution. It follows that the conditional probability that $X_j = 1$ is at most $\lambda + \tau$, and so $\mathbf{Exp}[X_i \cdot X_j] \leqslant \lambda(\lambda + \tau)$.

Thus, we have $\mathbf{Var}[X] \leqslant t\lambda + t^2\lambda(\lambda + \tau) - (t\lambda)^2 \leqslant t\lambda + t^2\lambda\tau = t\lambda(1 + t\tau) \leqslant 2t\lambda$ (since $t\tau \leqslant 1$ for our choice of t and τ), and so we get by Eq. (1) that

$$\mathbf{Pr}_{B \in R, A \in \mathbf{G}}\left[\left|\mathbf{Pr}_{i \in [t]}[(AS_i)(B) \in L'] - \lambda\right| \geqslant \lambda/3\right] \leqslant 18/(t\lambda).$$

Let $p = 18/(t\lambda) \leqslant 18/(t\alpha)$. By averaging, for all but at most $\sqrt{p}$ fraction of B's, it is the case that for all but at most $\sqrt{p}$ fraction of $A \in \mathbf{G}$, the fraction of subspaces $(AS_i)(B) \in L'$ is between $2\lambda/3$ and $4\lambda/3$.

Finally, for a given $B \in R$, the probability that a random d_l-dimensional linear subspace of B falls into L' is $\mathbf{Exp}_{A\in\mathbf{G}}\left[\mathbf{Pr}_{i\in[t]}[(AS_i)(B) \in L']\right]$. By the above, for all but $\sqrt{p}$ fraction of all $B \in R$, this average is at least $(1-\sqrt{p})(2/3)\lambda \geqslant \lambda/2$, and at most $(4/3)\lambda + \sqrt{p} \leqslant (3/2)\lambda$, since $\sqrt{p} < \lambda/10$ (by our assumption that $\lambda^{3/2}q^{(c-1)d_l/2} > 10$). □

2.4 Some properties of samplers

We will need some properties of samplers. Imagine the following setup. We take a bipartite graph $G = (L \cup R, E)$, choose a subset $L' \subseteq L$ of its left vertices of measure λ, and a subset $R' \subseteq R$ of its right vertices of measure ρ. Then we define the following distribution on vertices in L: Pick a uniformly random $r \in R'$, and output its uniformly random neighbor $\ell \in N(r)$. Clearly, if $\rho = 1$, we get the uniform distribution on L (since G is bi-regular), and so we hit the set L' with probability λ. The next lemma shows that, for sampler graphs G, the described distribution will hit L' with probability close to λ even for $\rho < 1$, provided that λ and ρ are sufficiently large.

Lemma 2.4. *Let $G = G(L, R)$ be any (α, β)-sampler. Let $0 \leqslant \lambda, \rho \leqslant 1$ be any values such that $\lambda \geqslant \alpha$ and $\lambda\rho/10 \geqslant \beta$. For any subset $L' \subseteq L$ of measure λ and any subset $R' \subseteq R$ of measure ρ, we have*

$$\left|\mathbf{Pr}_{r\in R',\ell\in N(r)}[\ell \in L'] - \lambda\right| \leqslant (2/3)\lambda.$$

Proof. The left-hand side of the required inequality is at most $\mathbf{Exp}_{r\in R'}\left[\left|\mathbf{Pr}_{\ell\in N(r)}[\ell \in L'] - \lambda\right|\right]$. By the definition of a sampler, we get for all but at most β/ρ fraction of vertices $r \in R'$ that $|\mathbf{Pr}_{\ell\in N(r)}[\ell \in L'] - \lambda| \leqslant \lambda/2$. So the overall expectation over $r \in R'$ is at most $\lambda/2 + \beta/\rho \leqslant \lambda/2 + \lambda/10$. □

Our definition of sampler graphs is asymmetric: every large set of *left* vertices is required to be sampled with approximately correct frequency by the neighborhood of almost every *right* vertex. The next lemma shows that a sampler graph actually enjoys a similar property for large sets of *right* vertices with respect to the neighborhoods of *left* vertices.

Lemma 2.5. *Let $G = G(L, R)$ be any (α, β)-sampler. Let $R' \subseteq R$ be any subset of measure ρ, and let $\lambda = \max\{\alpha, 10\beta/\rho\}$. Then for all but at most 2λ fraction of vertices $\ell \in L$, we have*

$$\left|\mathbf{Pr}_{r\in N(\ell)}[r \in R'] - \rho\right| \leqslant (2/3)\rho.$$

Proof. Let $Bad_1 \subseteq L$ be the subset of all those vertices $\ell \in L$ where $\mathbf{Pr}_{r\in N(\ell)}[r \in R'] > (5/3)\rho$, and let $Bad_2 \subseteq L$ be the subset of those vertices $\ell \in L$ where $\mathbf{Pr}_{r\in N(\ell)}[r \in R'] < (1/3)\rho$. We will argue that both Bad_1 and Bad_2 have measures less than λ.

If Bad_1 has measure at least λ, let us take a subset Bad_1' of Bad_1 of measure exactly λ. Consider picking a random edge in G. By the definition of Bad_1', the probability of picking an edge between Bad_1' and R' is greater than $\lambda(5/3)\rho$. On the other hand, by the definition of a sampler, this probability is at most $\rho(\lambda + \lambda/2) + \beta \leqslant \lambda\rho + 0.51\lambda\rho$. This contradiction shows that the measure of Bad_1 is less than λ.

Similarly, if Bad_2 has measure at least λ, we set Bad_2' to be its subset of measure exactly λ. The probability that a random edge is between Bad_2' and R' is less than $(1/3)\lambda\rho$. But, by the definition of a sampler, it must be at least $(\rho - \beta)\lambda/2 \geqslant \lambda\rho/2 - \lambda\rho/20 \geqslant 0.45\lambda\rho$. Hence, we must have Bad_2 of measure less than λ as well. □

For sampler graphs in the Independent case, we can show a tighter version of Lemma 2.5, as follows.

Lemma 2.6. *Let $G = (L \cup R, E)$ be the bipartite inclusion graph where $L = \mathcal{U}$, and R is a collection of all k-subsets. Let $R' \subseteq R$ be any subset of measure ρ. For any constant $0 < \nu < 1$, we have that for all but at most $O((\log 1/\rho)/k)$[7] fraction of vertices $\ell \in L$, we have*

$$\left|\mathbf{Pr}_{r \in N(\ell)}[r \in R'] - \rho\right| \leqslant \nu\rho.$$

Proof. Let $Bad_1 = \{\ell \in L \mid \mathbf{Pr}_{r \in N(\ell)}[r \in R'] \geq (1+\nu)\rho\}$, and let $Bad_2 = \{\ell \in L \mid \mathbf{Pr}_{r \in N(\ell)}[r \in R'] \leq (1-\nu)\rho\}$. We will argue that the measures of Bad_1 and Bad_2 are $O((\log 1/\rho)/k)$.

We start with Bad_2. Let $\lambda_2 = |Bad_2|/|L|$ be the measure of Bad_2. Suppose that $\lambda_2 > C(\log 1/\rho)/k$ for a large constant C to be specified later. Then, by the Chernoff-Hoeffding bounds, the probability over $r \in R$ that r contains at most $(1-\rho/2)\lambda_2 k$ elements from Bad_2 is $\exp(-\Omega(\lambda_2 k)) < \rho^{-\Omega(C)} < \nu\rho/8$, for sufficiently large C.

By the definition of Bad_2, we have

$$\mathbf{Pr}_{\ell \in L, r \in N(\ell)}[\ell \in Bad_2 \;\&\; r \in R'] \leqslant \lambda_2(1-\nu)\rho, \tag{2}$$

where the probability is over first picking a vertex $\ell \in L$ and then picking its random neighbor $r \in N(\ell)$. Since our bipartite graph G is bi-regular, we can compute the same probability by first picking a vertex $r \in R$, and then picking its random neighbor $\ell \in N(r)$. However, if we pick $r \in R$ first, the probability that $r \in R'$ and that it has at least $(1-\nu/2)\lambda_2 k$ elements in Bad_2 (and hence at least $(1-\nu/2)\lambda_2$ conditional probability that $\ell \in Bad_2$) is at least $(1-\nu/8)\rho$ (by our calculations above). That is, we have

$$\mathbf{Pr}_{r \in R, \ell \in N(r)}[r \in R' \;\&\; \ell \in Bad_2] \geqslant (1-\nu/8)\rho(1-\nu/2)\lambda_2 > \lambda_2\rho(1-(5/8)\nu),$$

contradicting (2) above.

Similarly, let $\lambda_1 = |Bad_1|/|L|$ and assume $\lambda_1 > C(\log 1/\rho)/k$ for a large constant C to be specified later. By the Chernoff-Hoeffding bounds, the probability over $r \in R$ that r contains at least $(1+\nu/2)\lambda_1 k$ elements from Bad_1 is $\exp(-\Omega(\lambda_1 k)) < \rho^{-\Omega(C)} < \nu\rho/64$, for a sufficiently large C. Moreover, for $D > 2e$, the probability that r contains more than $D\lambda_1 k$ elements from Bad_1 is at most

$$(e/D)^{D\lambda_1 k} < \rho^{CD} < (\nu\rho/64)^D < (\nu\rho)/(64)^D,$$

for a sufficiently large C.

Similarly to the case of Bad_2, we consider the following probability:

$$\mathbf{Pr}_{\ell \in L, r \in N(\ell)}[\ell \in Bad_1 \;\&\; r \in R'], \tag{3}$$

and bound it in two different ways. If we first pick $\ell \in L$, and then pick a random $r \in N(\ell)$, we get by the definition of Bad_1 that the probability in (3) is at least $\lambda_1(1+\nu)\rho$.

To bound this probability from above, we consider the partitioning of the set R' into the following sets, based on the number of intersections with the set Bad_1:

- $R_0 = \{r \in R' \mid |r \cap Bad_1| \leqslant (1+\nu/2)\lambda_1 k\}$,
- $R_1 = \{r \in R' \mid (1+\nu/2)\lambda_1 k < |r \cap Bad_1| < 8\lambda_1 k\}$, and,

[7] Here the hidden constant only depends on ν.

- for each integer $d \geqslant 8$, the set $R_d = \{r \in R' \mid d\lambda_1 k \leqslant |r \cap Bad_1| < (d+1)\lambda_1 k\}$.

Computing the probability in (3) as the sum of the corresponding probabilities for $r \in R_0$, $r \in R_1$, and $r \in R_d$ for all integer $d \geqslant 8$, we can bound it by

$$\rho(1+\nu/2)\lambda_1 + (\nu\rho/64)8\lambda_1 + \sum_{d \geqslant 8}(\nu\rho)(d+1)\lambda_1/64^d = \rho\lambda_1(1+\nu/2+\nu/8+\nu\sum_{d\geqslant 8}(d+1)/64^d),$$

which is less than $\rho\lambda_1(1+(3/4)\nu)$. But this contradicts our earlier lower bound $\rho\lambda_1(1+\nu)$ on the same probability. □

Corollary 2.7. *Let $G = (L \cup R, E)$ be the bipartite inclusion graph where L is the collection of all k'-subsets of the universe $\mathcal{U}$, and R is the collection of all k-subsets of $\mathcal{U}$, for some $k' < k$. Let $R' \subseteq R$ be any subset of measure $\rho < 1/2$. For any constant $0 < \nu < 1$, we have that for all but at most $O((\log 1/\rho)/(k/k'))$ fraction of vertices $\ell \in L$, we have*

$$\left|\mathbf{Pr}_{r \in N(\ell)}[r \in R'] - \rho\right| \leqslant \nu\rho.$$

Proof. Think of each $r \in R$ as a k/k'-tuple of sets of size k', and apply Lemma 2.6. □

Lemma 2.6 can be also interpreted as the following "average-case" version of the Chernoff-Hoeffding bound, which, to the best of our knowledge, has not been explicitly stated before.

Lemma 2.8 ("Average-case" Chernoff-Hoeffding bound)**.** *Let $S \subseteq \mathcal{U}$ be any subset of measure λ. Let $0 < \nu < 1$ be any constant. Let R^- be any subset of k-tuples of $\mathcal{U}$ such that $\mathbf{Exp}_{r \in R^-}[|r \cap S|] < (1-\nu)\lambda k$, and let R^+ be any subset of k-tuples such that $\mathbf{Exp}_{r \in R^+}[|r \cap S|] > (1+\nu)\lambda k$. Then the measure of each R^- and R^+ is at most $e^{-\Omega(\lambda k)}$, where the hidden constant only depends on ν.*[8]

3 Analysis of the direct-product tests

Our proof of Theorem 1.1 (and Theorem 1.2) is done in three stages, as described next.

Stage I: Low probability consistency implies high probability conditional consistency. In this stage, we show that any function C that has non-negligible chance of passing the V-test has very high probability of being similarly consistent *on the subset of instances for which it has good conditional probability of passing.*

More precisely, we show (in Section 3.1) that if the test accepts with probability at least ϵ, then the collection of all k-sets has the following structure. There are many (close to $\epsilon/2$ fraction) of k-sets (A_0, B_0) (with A_0 of size k') such that $C(A_0, B_0)|_{A_0} = C(A_0, B)|_{A_0}$ for many (at least $\epsilon/2$ fraction) of $(k-k')$-sets B, and, moreover, almost every pair of overlapping sets of the form (A_0, E, D_1) and (A_0, E, D_2) (where $|E| = |A_0|$) has the property: if $C(A_0, E, D_1)|_{A_0} = C(A_0, E, D_2)|_{A_0}$, then it is also the case that $C(A_0, E, D_1)|_E$ and $C(A_0, E, D_2)|_E$ agree in almost all positions.

Definition 3.1 (Consistency)**.** The sets B satisfying $C(A_0, B_0)|_{A_0} = C(A_0, B)|_{A_0}$ are called *consistent* with (A_0, B_0); we denote by $Cons_{A_0,B_0}$ the collection of all such consistent B's.

Definition 3.2 (Goodness)**.** We call (A_0, B_0) *good* if the collection $Cons_{A_0,B_0}$ has measure at least $\epsilon/2$.

[8] We assume here that, *on average*, the k-tuples in set R^- (resp., R^+) have too few (resp., too many) elements from a given set S. In contrast, the standard Chernoff-Hoeffding bound assumes that this happens for *each* k-tuple.

Definition 3.3 (Excellence)**.** We call (A_0, B_0) (α, γ)-*excellent* if it is good and, moreover,

$$\mathbf{Pr}_{E,D_1,D_2}[(E, D_i) \in Cons_{A_0,B_0},\ i = 1, 2,\ \ \& \ C(A_0, E, D_1)|_E \overset{>\alpha}{\neq} C(A_0, E, D_2)|_E] \leqslant \gamma,$$

where $|E| = |A_0| = k'$. (Think of $\alpha = \text{poly}(1/k)$ and $\gamma = \text{poly}(\epsilon)$.)

In this terminology, we show that there are at least about $\epsilon/2$ excellent k-sets (A_0, B_0). Note that for every excellent k-set (A_0, B_0), the $(k - k')$-sets $B \in Cons_{A_0,B_0}$ enjoy a very strong consistency property: *almost all* pairs of overlapping sets $B_1 = (E, D_1)$ and $B_2 = (E, D_2)$ from $Cons$ are such that $C(A_0, B_1)|_E$ and $C(A_0, B_2)|_E$ are almost identical.

Stage II: Unique decoding on a subset. Next, we show that we can do unique decoding on any subset such as $Cons_{A_0,B_0}$ above, where there is very high conditional probability of consistency. We can think of this as unique decoding of the direct product code where there are two types of noise: a very high number of erasures, and in addition a small number of values changed.

In Section 3.2, we use the strong consistency property of overlapping sets from $Cons_{A_0,B_0}$ (for an excellent set (A_0, B_0)) to show that there is a function g such that C computes the (approximate) direct product of g over almost all k-tuples $\{(A_0, B) \mid B \in Cons_{A_0,B_0}\}$. That is, there is a function g that is locally a direct-product function for C restricted to k-sets (A_0, B) for $B \in Cons_{A_0,B_0}$. This function g is defined very naturally as the plurality function: on input x, the value $g(x)$ is the most frequent value among the outputs of $C(A_0, B)$, over all $B \in Cons_{A_0,B_0}$ which contain x. (This is similar to the results in [FK00, DG08], but our proof techniques are different and yield better parameters.)

Stage III: Local decoding to global decoding. So far, the analysis used only the V-test, and showed that conditioned on being likely to pass the test, the answers to the first two oracle queries (A_0, B_0) and (A_0, B_1) are likely to be (almost) of the form: $g_{A_0}(B)$, a direct product for some function that depends only on A_0. Note that the counterexamples from [DG08] for the V-test have exactly this form, and show that, for $\epsilon < 1/k$, it is possible to have the above property, yet have very different functions g_A depending on the set A. The third query is meant to eliminate this possibility.

In Section 3.3, we use the third query (A_1, B_1) to argue that the same function g from the previous stage is actually also a *global* direct-product function for C on at least close to ϵ fraction of all possible k-sets.

Note that this third query is needed only if the acceptance probability $\epsilon < 1/k$. For the case of $\epsilon > \text{poly}(1/k)$ (more precisely, for $\epsilon \gg \sqrt{k'/k}$), we show in Section 3.4 that the two queries of the V-test alone suffice, thereby re-proving the result of [DG08].

Remark 3.4. It is easy to check that all results in the present section continue to hold also for the case of *randomized* oracle C (which supposedly computes some direct product function); the only change needed is to add internal randomness of C to all relevant probability expressions. However, for simplicity of notation, we will assume below that C is a deterministic oracle.[9]

3.1 Excellence

Using arguments similar to those in [IJKW08], we get the following.

Lemma 3.5. *Assume that* $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0}] \geqslant \epsilon$. *Then a random* (A_0, B_0) *is good with probability at least* $\epsilon/2$.

[9]The direct product analysis for a randomized oracle C will be used in Section 6 for the 2-query PCP construction.

Proof. Let $\mathcal{E}(A_0, B_0)$ be the event that (A_0, B_0) is not good, i.e., that $\mathbf{Pr}_{B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0}] < \epsilon/2$. Observe that $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0} \ \& \ \mathcal{E}(A_0, B_0)] \leqslant \mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0} \mid \mathcal{E}(A_0, B_0)] < \epsilon/2$. Hence, $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0} \ \& \ \neg\mathcal{E}(A_0, B_0)]$ is equal to

$$\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0}] - \mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0} \ \& \ \mathcal{E}(A_0, B_0)],$$

which is at least $\epsilon - \epsilon/2 = \epsilon/2$. Hence the event $\mathcal{E}(A_0, B_0)$ does not happen with probability at least $\epsilon/2$, as required. □

Lemma 3.6. $\mathbf{Pr}_{A_0,B_0}[(A_0, B_0)$ *is good but not* (α, γ)*-excellent*$] < \gamma'/\gamma$*, where* $\gamma' = e^{-\Omega(\alpha k')}$.

Proof. Set $\alpha' = \alpha/2$. The event in the statement of the lemma is the following event $\mathcal{E}_1(A_0, B_0)$: (A_0, B_0) is good but

$$\mathbf{Pr}_{E,D_1,D_2}[(E, D_i) \in Cons_{A_0,B_0},\ i = 1, 2 \ \& \ C(A_0, E, D_1)|_{A_0 \cup E} \overset{>\alpha'}{\neq} C(A_0, E, D_2)|_{A_0 \cup E}] > \gamma;$$

note that we allow α' errors in the set $A_0 \cup E$ of size $2|E|$, which for $(E, D_i) \in Cons$, $i = 1, 2$, means at most $2\alpha' = \alpha$ fraction of errors in the set E, as needed in the definition of (α, γ)-excellence.

Let $\mathcal{E}_2(A_0, B_0, E, D_1, D_2)$ be the event that (A_0, B_0) is good, $(E, D_i) \in Cons_{A_0,B_0}$ for $i = 1, 2$, and $C(A_0, E, D_1)|_{A_0 \cup E} \overset{>\alpha'}{\neq} C(A_0, E, D_2)|_{A_0 \cup E}$. Denote the set $A_0 \cup E$ by A'. The random choices of event $\mathcal{E}_2$ can be equivalently made in the following order: pick A', D_1, and D_2; pick A_0 as a random subset of A', setting $E = A' \setminus A_0$; pick random B_0. Condition on any (A', D_1) and (A', D_2) such that $C(A', D_1)_{A'} \overset{>\alpha'}{\neq} C(A', D_2)|_{A'}$. By the Chernoff-Hoeffding bound, a random k'-subset A_0 of A' will completely miss the inconsistent elements with probability at most $\gamma' = e^{-\Omega(\alpha' k')}$. If A_0 contains such inconsistent positions, then it cannot be the case that both $(E, D_1) \in Cons_{A_0,B_0}$ and $(E, D_2) \in Cons_{A_0,B_0}$. Hence, $\mathbf{Pr}[\mathcal{E}_2] \leqslant \gamma'$.

We have $\mathbf{Pr}[\mathcal{E}_2 \mid \mathcal{E}_1] > \gamma$. On the other hand, $\mathbf{Pr}[\mathcal{E}_2 \mid \mathcal{E}_1] = \mathbf{Pr}[\mathcal{E}_1 \ \& \ \mathcal{E}_2]/\mathbf{Pr}[\mathcal{E}_1] < \gamma'/\mathbf{Pr}[\mathcal{E}_1]$. So, we obtain that $\mathbf{Pr}[\mathcal{E}_1] < \gamma'/\gamma$, as required. □

As an immediate corollary of Lemmas 3.5 and 3.6, we get the following.

Corollary 3.7. *If* $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0, B_0)|_{A_0} = C(A_0, B_1)|_{A_0}] \geqslant \epsilon$*, then a random good set* (A_0, B_0) *is* (α, γ)*-excellent with probability at least* $1 - \epsilon^2$*, for* α *and* γ *such that* $\alpha k' > \Omega(\log 1/(\gamma\epsilon^3))$.

3.2 Excellence implies local agreement

Let us focus on $Cons = Cons_{A_0,B_0}$ for some fixed (α, γ)-excellent (A_0, B_0). Define the function g as follows: for every $x \in \mathcal{U} \setminus A_0$, set $g(x) = Plurality_{B \in Cons:\, x \in B}\ C(A_0, B)|_x$; if there is no $B \in Cons$ such that $x \in B$, then we set $g(x)$ to some default value, say 0.

Lemma 3.8. *There are fewer than* $\nu = O(\gamma/\epsilon^2) < \epsilon$ *fraction of sets* $B \in Cons$ *such that* $C(A_0, B)|_x \neq g(x)$ *for more than* $\beta = 40\alpha$ *fraction of* $x \in B$*, where* $\alpha > \Omega((\ln 1/\epsilon)/(k/k'))$.

We first give an outline of the proof of Lemma 3.8. For the sake of contradiction, suppose that $\mathbf{Pr}_{B \in Cons}[C(A_0, B)|_B \overset{>\beta}{\neq} g(B)] > \nu$, where $g(B)$ denotes the $|B|$-tuple of values of the direct product of g on the input set B. This means that

$$\mathbf{Pr}_{B \subseteq \mathcal{U} \setminus A_0}[B \in Cons \ \& \ C(A_0, B)|_B \overset{>\beta}{\neq} g(B)] > \nu', \tag{4}$$

for $\nu' > \nu\epsilon/2$ (since $Cons$ has measure at least $\epsilon/2$ by the definition of goodness of (A_0, B_0)).

Imagine choosing a random subset E of B. By Chernoff, we get that with probability close to 1, the set E has close to β fraction of inputs $x \in E$ where $C(A_0, B)|_x \neq g(x)$. Let $E' \subset E$ be the set of those $x \in E$ where C and g disagree.

On the other hand, using the definition of g as the plurality function as well as some basic sampling lemmas, we will show that, for almost every such random subset E of B and for the subset $E' \subseteq E$ defined as above, there are an $\Omega(\epsilon)$ fraction of $(k-k')$-sets B' such that $B' \in Cons$ and $C(A_0, B')|_{E'}$ agrees with $g(E')$ in $\Omega(1)$-fraction of positions.

Note that these two facts imply that $C(A_0, B)|_{E'}$ and $C(A_0, B')_{E'}$ disagree in a constant fraction of positions in E'. Since E' has size close to $\beta|E|$, we get that $C(A_0, B)|_E$ and $C(A_0, B')_E$ disagree in $\Omega(\beta)$ fraction of positions. This implies that one can pick, with non-negligible probability, a pair of sets B and B' with overlap E such that $B, B' \in Cons$ and $C(A_0, B)|_E$ and $C(A_0, B')|_E$ disagree in many positions, contradicting the excellence property of (A_0, B_0).

We provide the detailed proof next. We abstract away some of the parameters in the statement of Lemma 3.8, and re-state it as Lemma 3.10 below. Here, we prove the result for the Boolean case; in Section 5, we reduce the general case to the Boolean case.

Definition 3.9. Let $Cons$ be a subset of $\mathcal{U}^k$ of measure at least ϵ. Let C' be a function from $Cons$ to $\mathcal{R}^k$. We say C' is (α, γ)-*excellent with respect to* $Cons$ if the following holds: Pick $E \subset \mathcal{U}$ of size k', $D_1, D_2 \subset \mathcal{U}$ of size $k - k'$ independently at random. Then the probability that $E \cup D_1 \in Cons, E \cup D_2 \in Cons$ and $C'(E \cup D_1)|_E \overset{>\alpha}{\neq} C'(E \cup D_2)|_E$ is at most γ.[10]

Define the function g as before. That is, for every $x \in \mathcal{U}$, set $g(x) = Plurality_{B \in Cons:\, x \in B}\ C'(B)|_x$; if there is no $B \in Cons$ such that $x \in B$, then we set $g(x)$ to some default value, say 0.

Lemma 3.10. *Assume $\mathcal{R} = \{0, 1\}$. If C' is (α, γ)-excellent with respect to Cons, then there are fewer than $\nu = O(\gamma/\epsilon^2) < \epsilon$ fraction of sets $B \in Cons$ such that $C'(B)|_x \neq g(x)$ for more than $\beta = 40\alpha$ fraction of $x \in B$, where $\alpha > \Omega((\ln 1/\epsilon)/(k/k'))$.*

We will later prove the same lemma without the assumption that $\mathcal{R}$ is Boolean, but with a slightly worse value of β; see Section 5 below.

Towards a contradiction, suppose that $\mathbf{Pr}_{B \in Cons}[C'(B) \overset{>\beta}{\neq} g(B)] > \nu$, where $g(B)$ denotes the $|B|$-tuple of values of the direct product of g on the input set B. This means that

$$\mathbf{Pr}_{B \subseteq \mathcal{U}}[B \in Cons \ \&\ C'(B) \overset{>\beta}{\neq} g(B)] > \nu', \tag{5}$$

for $\nu' > \nu\epsilon$ (since $Cons$ has measure at least ϵ).

We will need the following notation. For each $x \in \mathcal{U}$, we denote by $\mathcal{B}_x$ the collection of all sets B that contain x, and let $Cons_x = Cons \cap \mathcal{B}_x$. Analogously, for each k'-subset $E \subset \mathcal{U}$, we denote by $\mathcal{B}_E$ the collection of all sets B that contain E, and let $Cons_E = Cons \cap \mathcal{B}_E$.

First we show that for almost all x, the measure of $Cons_x$ in $\mathcal{B}_x$ is large.

Claim 3.11. *For all but at most $O((\ln 1/\epsilon)/k)$ fraction of inputs $x \in \mathcal{U}$, we have $|Cons_x|/|\mathcal{B}_x| \geqslant \epsilon/6$.*

Proof. Apply Lemma 2.6. □

[10] We point out to the careful reader the following change in notation: before we had $Cons$ of measure $\epsilon/2$; k was $k - k'$; and $\mathcal{U}$ was $\mathcal{U} \setminus A_0$.

Claim 3.12. *Let x be any input such that $Cons_x$ has measure at least $\epsilon/6$ in $\mathcal{B}_x$. Then for all but at most $O((\ln 1/\epsilon)/(k/k'))$ fraction of k'-sets E containing x, we get that*

$$\mathbf{Pr}_{B \in Cons_E}[C'(B)|_x = g(x)] \geqslant 1/10.$$

Proof. Let $\mathcal{S}$ be a collection of all $(k'-1)$-size subsets E_x of $\mathcal{U}$, and let $\mathcal{T}$ be a collection of all $(k-1)$-size subsets B_x of $\mathcal{U}$. By assumption, we know that the measure μ of those sets B_x such that $B_x \cup \{x\} \in Cons$ is at least $\epsilon/6$. Let Q denote the set of all such sets B_x.

Let Q' be the subset of all those sets $B_x \in Q$ where $C(B_x \cup x)|_x = g(x)$. Let μ' be the measure of this Q' in $\mathcal{B}_x$. By the definition of g, we know that $\mu'/\mu \geqslant 1/2$, and so $\mu' \geqslant \epsilon/12$; here we use the assumption that g is a Boolean function.

Let $t = \lfloor |B_x|/|E_x| \rfloor \approx k/k' \approx \sqrt{k}$. By Corollary 2.7, we get that all but at most $\delta \leqslant O((\ln 1/\epsilon)/t)$ fraction of subsets E_x are such that, among the sets B_x containing E_x, the measure of those B_x that fall into Q is between $\mu/3$ and $5\mu/3$. Simultaneously, the measure of those $B_x \supset E_x$ that fall into Q' is between $\mu'/3$ and $5\mu'/3$, for all but at most δ fraction of subsets E_x. Hence, for at least $1-2\delta$ fraction of sets E_x, $\mathbf{Pr}_{B_x : E_x \subset B_x}[C'(B_x \cup x)|_x = g(x) \mid B_x \cup \{x\} \in Cons] \geqslant (\mu'/3)/(5\mu/3) \geqslant 1/10$, as required. □

Claim 3.13. *For $\delta = O((\ln 1/\epsilon)/(k/k'))$,*

$$\mathbf{Pr}_{E, x \in E}[\mathbf{Pr}_{B \in Cons_E}[C'(B)|_x = g(x)] \geqslant 1/10] \geqslant 1 - 2\delta.$$

Proof. The distribution $(E, x \in E)$ is the same as $(x, E \ni x)$. By Claim 3.11, we know that all but at most $O((\ln 1/\epsilon)/k)$ of x are such that $Cons_x$ is large. For each of these x, we get by Claim 3.12 that all but $O((\ln 1/\epsilon)/(k/k'))$ of E's will satisfy the event in the statement of the present claim. So over random choices of x and $E \ni x$, the required event occurs with probability at least $1 - O((\ln 1/\epsilon)/(k/k'))$. □

By a simple averaging argument, we get from Claim 3.13 the following corollary.

Claim 3.14. *Let $\delta = O((\ln 1/\epsilon)/(k/k'))$ be as in Claim 3.13, let $\delta' = 10\delta$, and let $\delta'' = 1/10$ (so that $\delta = \delta'\delta''$). For at least $1-\delta''$ fraction of sets E, we have that, for at least $1-\delta'$ fraction of inputs $x \in E$, $\mathbf{Pr}_{B \in Cons_E}[C'(B)|_x = g(x)] \geqslant 1/10]$.*

Finally, we will need the following analogue of Claim 3.11.

Claim 3.15. *For all but at most $O((\ln 1/\epsilon)/(k/k'))$ fraction of k'-subsets $E \subset \mathcal{U} \setminus A_0$, we have $|Cons_E|/|\mathcal{B}_E| \geqslant \epsilon/6$.*

Proof. Apply Corollary 2.7. □

We now give the proof of Lemma 3.10.

Proof of Lemma 3.10. Let $\delta' = 10\delta$ and $\delta'' = 1/10$, for the δ in Claim 3.14. We get by Claims 3.15 and 3.14 that, for at least $0.3 - o(1)$ fraction of uniformly random subsets E,

1. the fraction of sets $B' \supset E$ that fall into $Cons$ is at least $\epsilon/6$, and
2. for all but δ' fraction of inputs $x \in E$, $\mathbf{Pr}_{B' \in Cons_E}[C'(B')|_x = g(x)] \geqslant 1/10$.

Now consider the following distribution of subsets E: pick a random k-subset B satisfying the event of Eq. (5), and then pick a random k'-subset E of B. By Lemmas 2.3 and 2.4, we conclude that when E is sampled according to this distribution, we get with probability at least 0.29 a set E such that both conditions (1) and (2) above still hold.

For sets B and $E \subset B$, we denote by $E' \subseteq E$ the subset of those $x \in E$ where $C'(B)|_x \neq g(x)$. For every B satisfying the event of Eq. (5), we get by Chernoff-Hoeffding that almost all[11] subsets $E \subset B$ are such that $|E'| > (0.9\beta)|E|$. Combining this with our earlier argument, we get that for a random k-subset B satisfying the event of Eq. (5), if we pick a random subset $E \subset B$, we get with probability at least $0.29 - o(1) > 1/4$, a subset E such that conditions (1) and (2) above hold, and additionally, $|E'| > (0.9\beta)|E|$.

Fix any set E that satisfies the three conditions stated above. Let $E' \subset E$ be as above. Let $E'' \subseteq E'$ be the subset of those inputs $x \in E'$ where $\mathbf{Pr}_{B' \in Cons}[C'(B')|_x = g(x)] \geqslant 1/10$. By condition (2), we get that $|E''| \geqslant |E|(0.9\beta - \delta')$, which can be made at least $|E|\beta/2$ by choosing β sufficiently larger than δ' (as assumed in the statement of the lemma).

Thus, for every $x \in E''$, there are at least $1/10$ fraction of sets $B' \in Cons_E$ such that $C'(B')|_x = g(x) \neq C'(B)|_x$. By averaging, for at least $1/20$ fraction of $B' \in Cons_E$, we have $C'(B')|_x \neq C'(B)|_x$ for at least $1/20$ fraction of $x \in E''$. Since we also know that $|E''| > |E|\beta/2$, we get that

$$C'(B')|_E \overset{>\beta/40}{\neq} C'(B)|_E,$$

for at least $1/20$ fraction of $B' \in Cons_E$. By condition (1) on our fixed set E, we have that $Cons_E$ has measure at least $\epsilon/6$, and so

$$\mathbf{Pr}_{B':E\subset B'}[B' \in Cons \;\&\; C'(B')|_E \overset{>\beta/40}{\neq} C'(B)|_E] \geqslant \epsilon/120. \tag{6}$$

Since, for a random B conditioned on satisfying the event of Eq. (5), there are at least $1/4$ fraction of sets E such that Eq. (6) holds, we obtain

$$\mathbf{Pr}_{B,E\subset B,B'\supset E}[B' \in Cons \;\&\; C'(B')|_E \overset{>\beta/40}{\neq} C'(B)|_E \mid B \in Cons \;\&\; C'(B) \overset{>\beta}{\neq} g(B)] \geqslant \epsilon/480,$$

where the probability is over picking a random set B first, then picking its random k'-subset E, and finally picking a random set B' that contains E. Lifting the conditioning on the set B, we get

$$\mathbf{Pr}_{E,B\supset E,B'\supset E}[B' \in Cons \;\&\; C'(B')|_E \overset{>\beta/40}{\neq} C'(B)|_E \;\&\; B \in Cons] \geqslant \nu'\epsilon/480 > \nu\epsilon^2/960,$$

which contradicts the (α, γ)-excellence property for $\alpha = \beta/40$ and $\gamma = \nu\epsilon^2/960$. For $\gamma < \epsilon^3/960$, we get that $\nu < \epsilon$, as required. □

3.3 Local agreement implies global agreement

Here we prove the following lemma, which implies Theorem 1.1.

Lemma 3.16. *If the Z-test accepts with probability at least $\epsilon > e^{-\Omega(\alpha k')}$, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that for at least $\epsilon' = \epsilon/4$ fraction of all k-size sets S, the oracle $C(S)$ agrees with $g^k(S)$ in all but at most $\alpha' = 81\alpha$ fraction of inputs $x \in S$, where $k \geqslant \Omega(k'^2)$.*

[11] more precisely, all but at most $\exp(-\beta|E|)$ fraction

First we just sketch the argument, blurring over many details. Let (A_0, B_0) be randomly chosen in the first step of the Z-Test. If the test does not reject in step 2, we know that (A_0, B_0) is a good set, and moreover, by Corollary 3.7, it is an excellent set. By Lemma 3.8, we get that the oracle C on (almost all) k-sets (A_0, B), for $B \in Cons_{A_0,B_0}$, (mostly) agrees with the direct product of the majority function g (defined for $Cons_{A_0,B_0}$). We will argue that C will mostly agree with g^k also globally, on at least ϵ' fraction of all k-size sets S.

Consider picking sets B_1 and A_1 as follows: Pick a random k-set S, then randomly choose a subset $B_1 \subset S$, and set $A_1 = S \setminus B_1$; this choice of B_1 and A_1 is essentially equivalent to the way they are chosen by the Test. For the sake of contradiction, suppose that there are fewer than ϵ' sets S where C and g^k have agreement in more than $1 - \alpha'$ fraction of positions. Consider picking a random k-set S. If S is one of these ϵ' sets, then Test may accept, but this happens only with probability $\epsilon' < \epsilon$. So assume that S is a random k-set that contains more than α' fraction of inputs x where $C(S)|_x \neq g(x)$.

Pick a random subset B_1 of S of size $k - k'$; set $A_1 = S \setminus B_1$. If $B_1 \notin Cons_{A_0,B_0}$, Test will reject. Otherwise, by Lemma 3.8, we get that $g(B_1) = C(A_0, B_1)|_{B_1}$ on almost all inputs $x \in B_1$. At the same time, since $C(S) \overset{>\alpha'}{\neq} g(S)$, we get that with high probability $C(A_1, B_1)|_{B_1} \overset{>\alpha'/2}{\neq} g(B_1)$. But then $C(A_0, B_1)|_{B_1} \neq C(A_1, B_1)|_{B_1}$, and the Z-test rejects (in step 3). Thus, if there are few sets S where C and g^k have large agreement, the Z-test will accept with probability less than ϵ.

We now provide the detailed proof.

Proof of Lemma 3.16. Let (A_0, B_0) be randomly chosen in the first step of Test. Let κ be the measure of the set $Cons_{A_0,B_0}$. If this (A_0, B_0) is not good (i.e., if $\kappa < \epsilon/2$), then the set B_1 chosen in the second step of Test will be in $Cons_{A_0,B_0}$ with probability less than $\epsilon/2$. If this happens, then Test may accept, but only with the probability less than $\epsilon/2$.

Thus we need to analyze the case where (A_0, B_0) is a random good pair, and so $\kappa \geqslant \epsilon/2$. By Corollary 3.7, all but at most ϵ^2 of good pairs (A_0, B_0) are excellent. If our chosen good pair(A_0, B_0) is not excellent, then Test may accept, but this happens with probability at most $\epsilon^2 < \epsilon$.

We are left with the case where our chosen pair (A_0, B_0) is (α, γ)-excellent. For this pair, we define the function g as the majority function over sets in $Cons_{A_0,B_0}$. By Lemma 3.8, we know that C mostly agrees with the direct product of g on almost all k-sets (A_0, B), where $B \in Cons_{A_0,B_0}$. We will argue that C will mostly agree with g^k also globally, on at least ϵ' fraction of all k-size sets S.

Consider picking sets B_1 and A_1 as follows: Pick a random k-set S, then randomly choose a subset $B_1 \subset S$, and set $A_1 = S \setminus B_1$. This choice of B_1 and A_1 is essentially equivalent to the way they are chosen by the Test. The only difference is that the set B_1 chosen by the Test is disjoint from the set A_0, whereas in our new way of picking B_1 it may happen that B_1 intersects A_0. However, since B_1 is uniformly distributed, the probability that it intersects A_0 is negligible (less than $k^2/|\mathcal{U}|$). We will ignore this negligible amount, and think of this new choice of B_1 and A_1 as actually equivalent to the choices in the Test.

For the sake of contradiction, suppose that there are fewer than ϵ' sets S where C and g^k have agreement in more than $1 - \alpha'$ fraction of positions. Consider picking a random k-set S. If S is one of these ϵ' sets, then Test may accept, but this happens only with probability $\epsilon' < \epsilon$. So we may assume that S is a random k-set that contains more than α' fraction of inputs x where $C(S)|_x \neq g(x)$. Note that the distribution of such sets S is at most ϵ' far in statistical distance from the completely uniform distribution over k-sets S.

Pick a random subset B_1 of S of size $k - k'$; set $A_1 = S \setminus B_1$. If $B_1 \notin Cons_{A_0,B_0}$, Test will reject. So the probability that Test accepts is at most the probability that Test accepts conditioned

on $B_1 \in Cons_{A_0,B_0}$, times the probability that $B_1 \in Cons_{A_0,B_0}$.

By Lemma 3.8, all but at most $\nu = O(\gamma/\epsilon^2)$ fraction of sets $B \in Cons_{A_0,B_0}$ are such that $g(B)$ agrees with $C(A_0,B)|_B$ in all but at most $\beta = 40\alpha$ fraction of inputs $x \in B$. Conditioned on choosing a $B_1 \in Cons_{A_0,B_0}$, the Test chooses one of these ν fraction of sets with probability at most $(\nu\kappa + \epsilon')/(\kappa - \epsilon')$. Indeed, for a uniformly random set S, a random subset $B_1 \subset S$ is uniformly distributed, and so hits the ν-fraction of $Cons_{A_0,B_0}$ with probability at most $\nu\kappa$. For the distribution of sets S that is ϵ'-far from uniform, this hitting probability may increase by at most ϵ'. On the other hand, the probability that $B_1 \in Cons_{A_0,B_0}$ is at least $\kappa - \epsilon'$. The claimed bound follows.

At the same time, since $C(S) \overset{>\alpha'}{\neq} g(S)$, we get by Chernoff-Hoeffding that, for all but at most $e^{-\Omega(\alpha' k)}$ of $(k-k')$-subsets B_1 of S,

$$C(S)|_{B_1} \overset{>\alpha'/2}{\neq} g(B_1). \tag{7}$$

Conditioning on $B_1 \in Cons_{A_0,B_0}$ for a random S means that this probability gets multiplied by at most $1/(\kappa - \epsilon')$.

So conditioned on $B_1 \in Cons_{A_0,B_0}$, the probability that $B_1 \subset S$ for a random S was chosen so that either (7) is violated or that $g(B_1) \overset{>\beta}{\neq} C(A_0,B_1)|_{B_1}$ is at most $\rho = (\nu\kappa + \epsilon' + e^{-\Omega(\alpha k)})/(\kappa - \epsilon')$. In this case, Test may accept, but only with probability at most $\rho \cdot \mathbf{Pr}[B_1 \in Cons_{A_0,B_0}] \leqslant \rho(\kappa + \epsilon')$. This probability is less than ϵ, since $(\kappa + \epsilon')/(\kappa - \epsilon') \leqslant 3$ for $\epsilon' = \epsilon/4$ and $\kappa \geqslant \epsilon/2$.

Finally, for B_1 that satisfies (7) and is such that $g(B_1) \overset{\geqslant 1-\beta}{=} C(A_0,B_1)|_{B_1}$, we get that

$$C(A_1,B_1)|_{B_1} \overset{>\alpha'/2-\beta}{\neq} C(A_0,B_1)|_{B_1}.$$

Since $\alpha'/2 > \beta$, Test will reject in this case.

Thus we have argued that, in all cases, the probability that Test accepts is strictly less than ϵ. Hence, the function g must be an approximate direct-product function for C on ϵ' fraction of all k-sets. □

3.4 Two queries suffice when $\epsilon > \mathrm{poly}(1/k)$

Here we give a simpler proof of the following result of [DG08]. The same argument also yields Theorem 1.3.

Theorem 3.17 ([DG08])**.** *If the V-test accepts with probability at least $\epsilon \gg \sqrt{k'/k}$, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that for at least $\epsilon' = \Omega(\epsilon^6)$ fraction of all k-size sets S, the oracle $C(S)$ agrees with $g^k(S)$ in all but at most $1/k^{\Omega(1)}$ fraction of inputs $x \in S$.*

We provide two proofs of this theorem. The first one (in Section 3.4.1 below) is a direct argument using our earlier analysis of the V-test. The second proof (in Section 3.4.2) is by a reduction: we show that if V-test accepts with inverse-polynomial probability ϵ, then (a certain version of) the Z-test also accepts with probability $\mathrm{poly}(\epsilon)$, and hence the conclusion follows by the analysis of (the version of) the Z-test. Both proofs are short and simple, and rely on similar techniques.

3.4.1 Proof of Theorem 3.17

Key to the proof of this theorem is the ability to show that, if the V-test accepts, the following "double-excellence" holds. For many k-subsets S, two random *disjoint* k'-subsets A_1, A_2 of S

are simultaneously excellent[12]. With such pairs it is possible to move from "local consistency" to "global consistency" without an additional query (which was needed for exponentially small success probability). Indeed, we derive the existence of such pairs from the relatively high success probability assumed here. Moreover, the counterexample of [DG08] for sublinear success precisely precludes such disjoint excellent pairs.

Claim 3.18. *Assume the V-test accepts with probability $\epsilon \gg \sqrt{k'/k}$. Consider the following random experiment: Pick disjoint random k'-subsets $A_1, A_2 \subset \mathcal{U}$; pick random $(k-k')$-subsets $B_1 \subset \mathcal{U} \setminus A_1$ and $B_2 \subset \mathcal{U} \setminus A_2$; pick random $(k-2k')$-subset $B \subset \mathcal{U} \setminus (A_1 \cup A_2)$. Let $B' = B \cup A_1$, and let $B'' = B \cup A_2$. Then $\mathbf{Pr}[(A_i, B_i)$ is excellent, $i = 1, 2$, & $B' \in Cons_{A_2,B_2}$ & $B'' \in Cons_{A_1,B_1}] \geqslant \Omega(\epsilon^5)$.*

Proof. The experiment in the statement of the claim is equivalent to the following experiment: Pick random k-subset $S \subset \mathcal{U}$, randomly partition S into $\ell = k/k'$ subsets of size k' each; pick two distinct random k'-subsets A_1 and A_2 in this partition of S; pick random B_1 and B_2; set $B = S \setminus (A_1 \cup A_2)$ (and, as before, set $B' = B \cup A_1$ and $B'' = B \cup A_2$).

By Lemma 3.5 and Corollary 3.7, we know that, for random S, partition of S, and subset $A \subset S$ in this partition, the probability that $(A, (S \setminus A))$ is (α, γ)-excellent is at least $\epsilon/2(1-\epsilon^2) \geqslant \epsilon/3$. By averaging, for at least $\epsilon/6$ fraction of random sets S and partitions of S, there will be at least $\epsilon/6$ fraction of random subsets $A \subset S$ (chosen according to the partition of S) such that $(A, (S \setminus A))$ is excellent. Conditioned on picking such S and a partition of S, the probability of picking two disjoint excellent subsets $A_1, A_2 \subset S$ is at least $\epsilon^2/36 - 1/\ell$ (where $1/\ell$ is the probability of picking the same subset of S twice). Hence, the overall probability that both (A_1, B'') and (A_2, B') are excellent is at least $\Omega(\epsilon(\epsilon^2 - 1/\ell))$, which is at least $\Omega(\epsilon^3)$, since $1/\ell \ll \epsilon^2$ by the assumption.

Finally, conditioned on both (A_1, B'') and (A_2, B') being excellent, we get that $B_1 \in Cons_{A_1,B''}$ with probability at least $\epsilon/2$ and, similarly, $B_2 \in Cons_{A_2,B''}$ with probability at least $\epsilon/2$. That is, with probability at least $\epsilon^2/4$ over random B_1 and B_2, we get that $B'' \in Cons_{A_1,B_1}$ and $B' \in Cons_{A_2,B_2}$. Lifting the conditioning, we get that, with probability $\Omega(\epsilon^5)$, both (A_1, B'') and (A_2, B') are excellent, and $B'' \in Cons_{A_1,B_1}$ and $B' \in Cons_{A_2,B_2}$. This implies the claim since, for $B'' \in Cons_{A_1,B_1}$, the pair (A_1, B'') is excellent iff so is the pair (A_1, B_1) (and similarly for B'). □

Using Claim 3.18 and Lemma 3.8, we get the following claim.

Claim 3.19. *Assume the V-test accepts with probability $\epsilon \gg \sqrt{k'/k}$. Let $A_1, A_2, B_1, B_2, B, B', B''$ be as in the random experiment of Claim 3.18. Let g_{A_i} be the plurality function over sets in $Cons_{A_i,B_i}$, for $i = 1, 2$. For $\gamma \ll \epsilon^7$, we have $\mathbf{Pr}[(A_i, B_i)$ is (α, γ)-excellent, $i = 1, 2$, & $g_{A_1}(B) \overset{\leqslant O(\alpha)}{\neq} g_{A_2}(B)] \geqslant \Omega(\epsilon^5)$.*

Proof. Let $\beta = 40\alpha$. Conditioned on (A_1, B_1) being (α, γ)-excellent and on B'' being a random set in $Cons_{A_1,B_1}$, we get by Lemma 3.8 that $g_{A_1}(B'') \overset{>\beta}{\neq} C(A_1, B'')|_{B''}$ for fewer than $\gamma/\epsilon^2 \ll \epsilon^5$ fraction of random $(k-k')$-subsets B''; similarly, for (A_2, B_2) and B'. Together with Claim 3.18, this implies that the following event happens with probability at least $\Omega(\epsilon^5)$:

$$(A_i, B_i) \text{ is } (\alpha, \gamma)\text{-excellent, } i = 1, 2,\ g_{A_1}(B'') \overset{\leqslant\beta}{\neq} C(A_1, B'')|_{B''},\ g_{A_2}(B') \overset{\leqslant\beta}{\neq} C(A_2, B')|_{B'}.$$

The latter two equalities imply that $g_{A_1}(B) \overset{\leqslant\beta'}{\neq} C(A_1, B'')|_B$ and $g_{A_2}(B) \overset{\leqslant\beta'}{\neq} C(A_2, B')|_B$, for $\beta' \leqslant \beta(1 + o(1))$. Since $C(A_1, B'') = C(A_2, B')$, we conclude that $g_{A_1}(B) \overset{\leqslant 2\beta'}{\neq} g_{A_2}(B)$. □

[12] more precisely, both $(A_1, S \setminus A_1), (A_2, S \setminus A_2)$ are excellent.

Claim 3.20. *For at least $\Omega(\epsilon^5)$ fraction of random (A_1, B_1) and (A_2, B_2), we have that (A_1, B_1) and (A_2, B_2) are excellent, and that $g_{A_1}(x) = g_{A_2}(x)$ on all but $O(\alpha)$ fraction of inputs $x \in \mathcal{U}$.*

Proof. By Claim 3.19 and averaging, we get that for at least $\Omega(\epsilon^5)$ fraction of random (A_1, B_1) and (A_2, B_2), it is the case that (A_i, B_i) is excellent, for $i = 1, 2$, and that $\mathbf{Pr}_B[g_{A_1}(B) \overset{\leqslant \alpha'}{\neq} g_{A_2}(B)] \geqslant \Omega(\epsilon^5)$, for some $\alpha' = O(\alpha)$. Fix any such (A_1, B_1) and (A_2, B_2). Suppose that $\mathbf{Pr}_{x \in \mathcal{U}}[g_{A_1}(x) \neq g_{A_2}(x)] > 2\alpha'$. Pick a random $B \subset \mathcal{U} \setminus (A_1 \cup A_2)$ of size $k - 2k'$. By Chernoff, the probability that $g_{A_1}(B) \overset{\leqslant \alpha'}{\neq} g_{A_2}(B)$ is less than $\nu = e^{-\Omega(\alpha'|B|)}$. This is a contradiction since $\nu \ll e^5$. □

Proof of Theorem 3.17. By Claim 3.19 and an averaging argument, we get that there are $\Omega(\epsilon^5)$ pairs (A_1, B_1) such that $\mathbf{Pr}_{A_2,B_2,B}[(A_2, B_2)$ is (α, γ)-excellent & $g_{A_1}(\mathcal{U}) \overset{\leqslant \alpha'}{\neq} g_{A_2}(\mathcal{U})] \geqslant \Omega(\epsilon^5)$, where A_2, B_2, B are chosen as in the random experiment of Claim 3.19, and $\alpha' = O(\alpha)$. Fix any such (A_1, B_1). We show that C is close to the direct product of g_{A_1} on poly(ϵ) fraction of k-sets $S \subset \mathcal{U}$.

Picking a random k-set S is equivalent to picking disjoint random subsets A_2 and E, of size k' each, B_2 of size $k - k'$, and B of size $k - 2k'$, and setting $S = B \cup A_2 \cup E$. Condition on the event that random (A_2, B_2) is excellent and g_{A_1} and g_{A_2} disagree on at most α' fraction of inputs in $\mathcal{U}$; this event happens with probability $\Omega(\epsilon^5)$. Further condition on the event that $(B \cup E) \in Cons_{A_2,B_2}$; this event happens with probability $\Omega(\epsilon)$ (given the previous conditioning on (A_2, B_2)).

Given these conditionings, we get by Lemma 3.8 that, with probability $1 - o(1)$, $g_{A_2}(B \cup E) = C(S)|_{B \cup E}$ in all but at most $O(\alpha)$ fraction of positions. By Chernoff, with probability $1 - \exp(\alpha|B|) \geqslant 1 - o(1)$, $g_{A_1}(B \cup E) = g_{A_2}(B \cup E)$ in all but at most $O(\alpha)$ fraction of positions. Hence, with probability $1 - o(1)$, $g_{A_1}(B \cup E) = C(S)|_{B \cup E}$ except for $O(\alpha)$ fraction of positions, and thus $g_{A_1}(S) = C(S)$ except for $O(\alpha k)$ positions (since $k'/k \leqslant O(\alpha)$). Lifting the conditionings, we get, for $\Omega(\epsilon^6)$ of random k-sets $S \subset \mathcal{U}$, that $g_{A_1}(S) = C(S)$ except for $O(\alpha k)$ positions. □

3.4.2 Alternative proof of Theorem 3.17

In this section, we give an alternative analysis of the V-test and derandomized V-test when $q > k^{-\alpha}$ for some absolute constant α.

We'll start with an alternative construction of a 3-query direct product test, the Z'-test; this Z'-test is sound for essentially the same reason as the Z-test defined earlier (for completeness, we give the proof in the Appendix). We then give a third test, the correlated twice-V test, cV^2. We show that the acceptance probability of the correlated twice-V test is at least q^2. Then we show that the acceptance probability of the Z'-test is at least that of the cV^2-test, less some polynomial in k. It then follows that the Z'-test accepts with probability at least $q^2 - \text{poly}(1/k)$. Thus, if $q \gg k^{-\alpha}$, then the Z'-test accepts with probability poly(q). Since the Z'-test is sound (even with inverse-exponential soundness), this implies that C is close to a direct product function. The analogous argument will also work fairly directly for the derandomized case.

Assume $q > k^{-1/4}/4$. The Z'-test is as follows:

Z′-Test:
1. Pick a random k-set $(A_0, B_0) \subseteq \mathcal{U}$, where $|A_0| = k' = q\sqrt{k}/2$.
2. Pick a random k-set $(A_1, B_1) \subseteq \mathcal{U}$, where $|A_1| = k'$.
3. Pick a random set $M \subseteq \mathcal{U} \setminus (A_0 \cup A_1)$ of size $k - 2k'$. If $C(A_0, B_0)|_{A_0} \neq C(A_0, M \cup A_1)|_{A_0}$ or if $C(A_1, B_1)|_{A_1} \neq C(A_1, M \cup A_0)|_{A_1}$ then reject; else, accept.

Pictorially, the Z'-test is given in Fig. 2 below. Note that Z'-test is more symmetric than the Z-test: the Z'-test consists of two identical V-tests "glued together".

The cV^2-test is:

cV^2-Test:
1. Pick a random k-set $N_0 \subseteq \mathcal{U}$.
2. Pick $A_0 \subseteq N_0$, $|A_0| = k'$, and $B_0 \subseteq \mathcal{U} \setminus A_0$. If $C(A_0, B_0)|_{A_0} \neq C(A_0, N_0 - A_0)|_{A_0}$ then reject, else continue.
3. Pick $A_1 \subseteq N_0$, $|A_1| = k'$, and $B_1 \subseteq \mathcal{U} \setminus A_1$.
If $C(A_1, B_1)|_{A_1} \neq C(A_1, N_0 - A_1)|_{A_1}$ then reject, else accept.

Assume C passes the V-test with probability $\epsilon \geq q$. Let ϵ_N be the conditional probability of passing the V-test given that $A_0 \cup B_0 = N$. Then $\mathbf{Exp}_{N \subseteq \mathcal{U}, |N|=k}[\epsilon_N] = \epsilon$, and the probability of passing the cV^2 test given that $N_0 = N$ is $(\epsilon_N)^2$. Thus, the probability of passing the cV^2 test overall is $\mathbf{Exp}_N[(\epsilon_N)^2] \geq (\mathbf{Exp}_N[\epsilon_N])^2 = \epsilon^2$, by Cauchy-Schwarz.

Note that, given that A_0 and A_1 are disjoint in both tests, the Z' test and cV^2 tests are identically distributed (setting $N = A_0 \cup A_1 \cup M$). Thus, the probability that C passes the Z' test is at least $\epsilon^2 - \mathbf{Pr}[(A_0 \cap A_1) \neq \varnothing] \geqslant \epsilon^2 - (k')^2/k = \epsilon^2 - q^2/4 \geq (3/4)q^2$. So Theorem 3.17 follows from the analysis of the Z'-test, given in the next theorem.

Theorem 3.21 (Analysis of the Z'-test)**.** *Assume C passes the Z'-test with probability p. Then there exists a function $G : \mathcal{U} \to \mathcal{R}$ so that with probability at least $\Omega(p^2)$ over k-sets S,*

$$G^k(S) \overset{\geq 1 - O((\log 1/p)k'/k)}{=} C(S).$$

The proof of Theorem 3.21 is very similar to (in fact, even simpler than) the analysis of the Z-test given in Lemma 3.16. For completeness, we prove Theorem 3.21 in the Appendix.

4 Derandomized DP testing: Proofs of Theorems 1.2 and 1.3

Here we prove Theorem 1.2. Our proof follows the structure of the proof for the Independent case from the previous section, but now relying on the *subspace* samplers from Lemma 2.3. We will fully prove Theorem 1.2. The proof of Theorem 1.3 is in fact simpler, using the first two parts of the proof of Theorem 1.2 and a simple additional property of the derandomized partitions.

4.1 Excellence

For a pair (A_0, B_0) (where A_0 and B_0 are linearly independent subspaces), we say that a subspace B (linearly independent from A_0) is *consistent with* (A_0, B_0) if $C(A_0 + B_0)|_{A_0} = C(A_0 + B)|_{A_0}$. Let $Cons_{A_0,B_0}$ denote the set of all such consistent subspaces B.

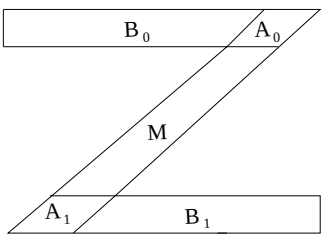

Figure 2: Z'-test.

As before, we say that (A_0, B_0) is *good* if $Cons_{A_0,B_0}$ has measure at least $\epsilon/2$. We call (A_0, B_0) (α, γ)-*excellent* if it is good and, moreover,

$$\mathbf{Pr}_{E,D_1,D_2}[(E+D_i) \in Cons_{A_0,B_0},\ i=1,2,\ \ \&\ C(A_0+E+D_1)|_{A_0+E} \overset{>\alpha}{\neq} C(A_0+E+D_2)|_{A_0+E}] \leqslant \gamma,$$

where E, D_1 and D_2 are linear subspaces such that E, D_1 and D_2 are linearly independent from A_0, and each D_i is linearly independent from E. The dimension of E is the same as that of A_0.

We get the following analogues of Lemmas 3.5 and 3.6.

Lemma 4.1. *Assume that* $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0 + B_0)|_{A_0} = C(A_0 + B_1)|_{A_0}] \geqslant \epsilon$. *Then a random* (A_0, B_0) *is good with probability at least* $\epsilon/2$.

Proof. The proof is exactly the same as that of Lemma 3.5. □

Lemma 4.2. $\mathbf{Pr}_{A_0,B_0}[(A_0, B_0)$ *is good but not* (α, γ)-*excellent*$] < \gamma'/\gamma$, *where* $\gamma' \leqslant O(q^2/(\alpha k'))$.

Proof. Set $\alpha' = \alpha/2$. We need to upperbound the probability of the event $\mathcal{E}_1(A_0, B_0)$ that (A_0, B_0) is good but

$$\mathbf{Pr}_{E,D_1,D_2}[(E+D_i) \in Cons_{A_0,B_0},\ i=1,2\ \ \&\ C(A_0+E+D_1)|_{A_0+E} \overset{>\alpha'}{\neq} C(A_0+E+D_2)|_{A_0+E}] > \gamma.$$

Let $\mathcal{E}_2(A_0, B_0, E, D_1, D_2)$ be the event that (A_0, B_0) is good, $(E + D_i) \in Cons_{A_0,B_0}$ for $i = 1, 2$, and $C(A_0 + E + D_1)|_{A_0+E} \overset{>\alpha'}{\neq} C(A_0 + E + D_2)|_{A_0+E}$. Denote the subspace $A_0 + E$ by A. The random choices of event $\mathcal{E}_2$ can be equivalently made in the following order: pick a subspace A, and subspaces D_1 and D_2 linearly independent from A; pick A_0 as a random subspace of A, setting E to be a random subspace of A that is linearly independent from A_0; pick random B_0 linearly independent from A. Condition on any (A, D_1) and (A, D_2) such that $C(A+D_1)|_A \overset{>\alpha'}{\neq} C(A+D_2)|_A$. A random subspace A_0 of A will completely miss the inconsistent elements in A with probability at most $\gamma' \leqslant O(q^2/(\alpha'|A_0|)) \leqslant \mathrm{poly}(1/k)$, by Lemma 2.2 (with $V_0 = \{0\}$, $V_1 = A$, and $W = A_0$). If A_0 contains such inconsistent positions, then it cannot be the case that both $(E+D_1) \in Cons_{A_0,B_0}$ and $(E + D_2) \in Cons_{A_0,B_0}$. Hence, $\mathbf{Pr}[\mathcal{E}_2] \leqslant \gamma'$.

We have $\mathbf{Pr}[\mathcal{E}_2 \mid \mathcal{E}_1] > \gamma$. On the other hand, $\mathbf{Pr}[\mathcal{E}_2 \mid \mathcal{E}_1] = \mathbf{Pr}[\mathcal{E}_1\ \&\ \mathcal{E}_2]/\mathbf{Pr}[\mathcal{E}_1] < \gamma'/\mathbf{Pr}[\mathcal{E}_1]$. So, we obtain that $\mathbf{Pr}[\mathcal{E}_1] < \gamma'/\gamma$, as required. □

As a consequence of these two lemmas, we get

Corollary 4.3. *Assume that* $\mathbf{Pr}_{A_0,B_0,B_1}[C(A_0 + B_0)|_{A_0} = C(A_0 + B_1)|_{A_0}] \geqslant \epsilon$. *Then we have*

$$\mathbf{Pr}_{A_0,B_0}[(A_0, B_0)\ \textit{is}\ (\alpha, \gamma)\textit{-excellent} \mid (A_0, B_0)\ \textit{is good}] \geqslant 1 - \epsilon^2,$$

where α *and* γ *are such that* $\alpha k' > \mathrm{poly}(\gamma\epsilon^3)$.

4.2 Excellence implies local agreement

Let us focus on $Cons_{A_0,B_0}$ for some fixed (α, γ)-excellent (A_0, B_0). For notational convenience, we will drop the subscript and simply write $Cons$. By the excellence property, we have the following:

$$\mathbf{Pr}_{E,D_1,D_2}[(E+D_i) \in Cons,\ i=1,2\ \ \&\ C(A_0+E+D_1)|_{A_0+E} \overset{>\alpha}{\neq} C(A_0+E+D_2)|_{A_0+E}] \leqslant \gamma, \quad (8)$$

where E is a d_0-dimensional subspace independent from A_0, and each D_i is a $(d-2d_0)$-dimensional subspace independent from the subspace $A_0 + E$.

Define the function g as follows. For every $x \in A_0$, set $g(x) = C(A_0+B_0)|_x$; for every $x \in \mathcal{U} \setminus A_0$, set

$$g(x) = Plurality_{B \in Cons:\ x \in A_0+B}\ \ C(A_0 + B)|_x; \tag{9}$$

if there is no $B \in Cons$ such that $x \in A_0 + B$, then we set $g(x)$ to some default value, say 0.

We will prove the following.

Lemma 4.4. *There are fewer than $\nu = O(\gamma/\epsilon^2) < \epsilon$ fraction of $B \in Cons$ such that $C(A_0+B)|_x \neq g(x)$ for more than $\beta = 40\alpha$ fraction of $x \in A_0 + B$.*

For the sake of contradiction, suppose that $\mathbf{Pr}_{B \in Cons}[C(A_0+B) \overset{>\beta}{\neq} g(A_0+B)] > \nu$. This implies that

$$\mathbf{Pr}_B[B \in Cons\ \&\ C(A_0 + B) \overset{>\beta}{\neq} g(A_0 + B)] > \nu', \tag{10}$$

for $\nu' > \nu\epsilon/2$ (since $Cons$ has measure at least $\epsilon/2$ by the definition of goodness of (A_0, B_0)).

We will follow the structure of the proof argument used to prove the analogous result for the Independent case of Lemma 3.8. For technical reasons, we will use different methods for sampling subspaces containing A_0 than those used in Eqs. (8)–(10), and the statement of Lemma 4.4 above. However, these alternative sampling methods will preserve the values of the probabilities of the corresponding events. We give the details in the following subsection.

4.2.1 Equivalent sampling procedures

Subsets of a universe $\mathcal{U}$ containing a fixed set A_0, are in 1-1 correspondence with subsets of the complement, $\mathcal{U} \setminus A_0$. A similar statement holds for *subspaces*, i.e., when $\mathcal{U}$ is a vector space, A_0 is a fixed subspace, and we consider subspaces containing A_0. Here we formally describe this correspondence.

For every linear subspace L of the universe $\mathcal{U} = \mathbb{F}_q^m$, we have a complementary subspace L' that is disjoint from L except for the common zero vector and such that $L + L' = \mathcal{U}$. In general, there are many subspaces complementary to the given subspace L. Among all such spaces, we can choose some canonical one, and denote by $L^\dagger$ this uniquely defined subspace complementary to L. (In the case of, say, real fields, we could simply take $L^\dagger$ to be the uniquely defined orthogonal complement of L.)

Let A_0 be a fixed linear subspace of $\mathcal{U} = \mathbb{F}_q^m$. For every subspace B linearly independent from A_0, there is a subspace $B_\perp \subseteq (A_0)^\dagger$ such that $A_0 + B = A_0 + B_\perp$. Indeed, one can take basis vectors of B, represent each of them as a sum $a + b$ for $a \in A_0$ and $b \in (A_0)^\dagger$, and take the corresponding vectors b as the basis for $B_\perp$. It is easy to see that $B_\perp$ is uniquely determined by the space $A_0 + B$; i.e., for any subspaces B', B'' of $(A_0)^\dagger$, if $A_0 + B' = A_0 + B''$ then $B' = B''$.

Let us call two subspaces B and B' equivalent if $A_0 + B = A_0 + B'$. All such equivalence classes are of the same size, and, by the above, they are in one-to-one correspondence with the subspaces in $(A_0)^\dagger$.

Let $\mathcal{E}(B)$ be any random event (of a random subspace B) which depends on the properties of the space $A_0 + B$ (rather than a particular representative of the equivalence class of B). Then the probability of the event $\mathcal{E}(B)$ for a uniformly chosen subspace B linearly independent from A_0 is equal to that for a uniformly chosen subspace B from the space $(A_0)^\dagger$.

For example, let $\mathcal{B}$ denote the set of all $(d - d_0)$-dimensional subspaces independent from A_0, and let $\mathcal{B}_\perp$ denote the set of all $(d - d_0)$-dimensional subspaces in $(A_0)^\dagger$. Recall that $Cons$ is the

subset of all those $B \in \mathcal{B}$ such that $C(A_0 + B_0)|_{A_0} = C(A_0 + B)|_{A_0}$. Define $Cons_\perp = Cons \cap \mathcal{B}_\perp$. We get that the measure of consistent subspaces remains the same when we sample subspaces from $(A_0)^\dagger$ (rather than from all linearly independent subspaces). That is, we have the following.

Claim 4.5. $|Cons|/|\mathcal{B}| = |Cons_\perp|/|\mathcal{B}_\perp|$.

For each $x \in \mathcal{U} \setminus A_0$, let $\mathcal{B}_x$ denote the collection of all subspaces B linearly independent from A_0 such that $x \in A_0 + B$. We denote by $Cons_x$ the subset of $Cons$ that consists of exactly those $B \in Cons$ such that $x \in A_0 + B$.

Each $x \in \mathcal{U} \setminus A_0$ can be uniquely represented as $x_\| + x_\perp$, where $x_\| \in A_0$ and $0 \neq x_\perp \in (A_0)^\dagger$. Let L_x denote the 1-dimensional linear subspace spanned by $x_\perp$. Let $(\mathcal{B}_x)_\perp$ denote the set of all $(d - d_0)$-dimensional subspaces from $(A_0)^\dagger$ that contain L_x. Let $(Cons_x)_\perp$ denote the subset of those $B \in (\mathcal{B}_x)_\perp$ that are in $Cons$. We get the following.

Claim 4.6. $|Cons_x|/|\mathcal{B}_x| = |(Cons_x)_\perp|/|(\mathcal{B}_x)_\perp|$.

For each subspace E linearly independent from A_0, we denote by $Cons_E$ the subset of $Cons$ that consists of exactly those $B \in Cons$ that contain E, and we denote by $\mathcal{B}_E$ the collection of all B linearly independent from A_0 such that $E \subseteq B$. Let $(\mathcal{B}_E)_\perp$ be the set of all subspaces from $(A_0)^\dagger$ that contain $E_\perp$, and let $(Cons_E)_\perp$ be the set of all those subspaces $B' \in (\mathcal{B}_E)_\perp$ that are in $Cons$. We have the following analogue of Claim 4.6.

Claim 4.7. $|Cons_E|/|\mathcal{B}_E| = |(Cons_E)_\perp|/|(\mathcal{B}_E)_\perp|$.

One can easily show that the probability in Eq. (8) remains the same when one samples the subspaces E, D_1, D_2 as follows: choose a uniform d_0-dimensional subspace E from $(A_0)^\dagger$, choose two independently random $(d-2d_0)$-dimensional subspaces D_1 and D_2 from the subspace $(A_0+E)^\dagger$. Similarly, one can change the sampling method in Eq. (10) (to subspaces $B \in \mathcal{B}_\perp$), without changing the probability value. Finally, the function g defined in Eq. (9) remains the same when one samples Bs from $(Cons_x)_\perp$ rather than from $Cons_x$.

4.2.2 Proof of Lemma 4.4

We now show the analogues of Claims 3.11–3.15. By Claims 4.6 and 4.7, it is sufficient for us to argue about the corresponding $\perp$-versions of the involved sets of subspaces. To simplify the notation, in the rest of this subsection we shall drop the subscript $\perp$ when denoting these versions.

Claim 4.8. *For all but at most $1/(\epsilon k^{1/5})$ fraction of inputs $x \in \mathcal{U} \setminus A_0$, we have $|Cons_x|/|\mathcal{B}_x| \geqslant \epsilon/6$.*

Proof. Consider the bipartite graph where the left vertices are labeled by all 1-dimensional linear subspaces from $(A_0)^\dagger$, and the right vertices are labeled by all linear D-dimensional subspaces from $(A_0)^\dagger$, for $D = d - d_0$. By Lemma 2.3, this graph is a (β, λ)-sampler, for $\lambda = O(1/\sqrt{\beta q^{D-2}})$.

We know that $Cons$ has measure $\mu \geqslant \epsilon/2$ among all the vertices on the right. Let H be the subset of all those vertices on the left with fewer than $\mu/3$ fraction of their neighbors falling into $Cons$. By Lemma 2.5, we can conclude that the measure of H is at most $1/(\epsilon q^{D/4})$, which is at most $1/(\epsilon k^{1/5})$ (by our choice of D and d_0).

Finally, since choosing a random $x \in \mathcal{U} \setminus A_0$ is equivalent to choosing a random vector in A_0, a random 1-dimensional subspace L from $(A_0)^\dagger$, and a random non-zero vector in L, we conclude that choosing a random x with $|Cons_x|/|\mathcal{B}_x| < \epsilon/6$ is at most the measure of H. □

Next we prove the following analogue of Claim 3.12.

Claim 4.9. *Let x be any input such that $Cons_x$ has measure at least $\epsilon/6$ in $\mathcal{B}_x$. Then for all but at most $O(1/(\epsilon k^{1/4}))$ fraction of linear subspaces E from $(A_0)^\dagger$ such that $x \in A_0 + E$, we get that*

$$\mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \geqslant 1/10.$$

Proof. Let $x = x_\| + x_\perp$ be the unique representation of x such that $x_\| \in A_0$ and $x_\perp \in (A_0)^\dagger$. Let L_x be the linear subspace spanned by the vector $x_\perp$.

A $(d - d_0)$-dimensional linear subspace $B \in \mathcal{B}_x$ is uniquely determined by a $(d - d_0 - 1)$-dimensional linear subspace B' from $(A_0 + L_x)^\dagger$ (with $B = B' + L_x$). Similarly, a linear d_0-dimensional subspace E (from $(A_0)^\dagger$) such that $x \in A_0 + E$ is uniquely determined by a $(d_0 - 1)$-dimensional linear subspace E' from $(A_0 + L_x)^\dagger$ (with $E = E' + L_x$).

Consider the bipartite inclusion graph with the left vertices labeled by all $(d_0 - 1)$-dimensional linear subspaces E', and the right vertices labeled by all $(d - d_0 - 1)$-dimensional linear subspaces B' (all from $(A_0 + L_x)^\dagger$). Let Q be the subset of those right vertices B' such that $B' + L_x \in Cons$. By the assumption, Q has measure $\mu \geqslant \epsilon/6$.

Let Q' be the subset of all those $B' \in Q$ where $C(A_0 + B' + L_x)|_x = g(x)$. Let μ' be the measure of this Q' in the set of right vertices. By the definition of g, we know that $\mu'/\mu \geqslant 1/2$, and so $\mu' \geqslant \epsilon/12$.

Let $c = \lfloor (d - d_0 - 1)/(d_0 - 1) \rfloor \approx 24$. By Lemmas 2.3 and 2.5, all but at most $\delta \leqslant 1/(\epsilon k^{1/4})$ fraction of subspaces E' are such that, among the subspaces B' containing E', the measure of those B' that fall into Q is between $\mu/3$ and $5\mu/3$. Simultaneously, the measure of those $B' \supset E'$ that fall into Q' is between $\mu'/3$ and $5\mu'/3$, for all but at most δ fraction of subsets E'. Hence, for at least $1 - 2\delta$ fraction of sets E', $\mathbf{Pr}_{B' : E' \subset B'}[C(A_0 + B' + L_x)|_x = g(x) \mid B' + L_x \in Cons] \geqslant (\mu'/3)/(5\mu/3) \geqslant 1/10$, as required. □

Claim 4.10. *For $\delta = O(1/(\epsilon k^{1/5}))$,*

$$\mathbf{Pr}_{E, x \in (A_0 + E) \setminus A_0}[\mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \geqslant 1/10] \geqslant 1 - \delta,$$

where E is a random d_0-dimensional linear subspace from $(A_0)^\dagger$.

Proof. The distribution $(E, x \in (A_0 + E) \setminus A_0)$ is the same as $(x \in \mathcal{U} \setminus A_0, E : x \in A_0 + E)$. By Claim 4.8, we know that all but at most $1/(\epsilon k^{1/5})$ of x are such that $Cons_x$ is large. For each of these x, we get by Claim 4.9 that all but $O(1/(\epsilon k^{1/4}))$ of Es will satisfy the event in the statement of the present claim. So over random choices of x and E such that $x \in A_0 + E$, the required event occurs with probability at least $1 - O(1/(\epsilon k^{1/5}))$. □

Claim 4.11. $\mathbf{Pr}_E[\forall x \in A_0 + E \quad \mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \geqslant 1/10] \geqslant 1 - \delta'$, *for* $\delta' = O(1/(\epsilon k^{3/25}))$.

Proof. The proof is by the union bound. Consider the set of all those E with at least one $x \in A_0 + E$ such that $\mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \leqslant 1/10$. Let μ be the measure of this set of Es. It follows that $\mathbf{Pr}_{E, x \in A_0 + E}[\mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \leqslant 1/10] \geqslant \mu/|A_0 + E| = \mu/q^{2d_0} = \mu/k^{2/25}$.

On the other hand, by Claim 4.10, we have that $\mathbf{Pr}_{E, x \in A_0 + E}[\mathbf{Pr}_{B \in Cons_E}[C(A_0 + B)|_x = g(x)] \leqslant 1/10] \leqslant \delta$, for $\delta = O(1/(\epsilon k^{1/5}))$. We conclude that $\mu \leqslant \delta k^{2/25} \leqslant O(1/(\epsilon k^{3/25}))$, as required. □

Finally, we will need the following analogue of Claim 4.8.

Claim 4.12. *For all but at most $1/(\epsilon k^{1/4})$ fraction of d_0-dimensional linear subspaces E from $(A_0)^\dagger$, we have $|Cons_E|/|\mathcal{B}_E| \geqslant \epsilon/6$.*

Proof. The proof is analogous to that of Claim 4.8, except we use the inclusion graph on vertices that are d_0-dimensional linear subspaces E and $(d-d_0)$-dimensional linear subspaces B, with both E and B from $(A_0)^\dagger$. □

We now give the proof of Lemma 4.4.

Proof of Lemma 4.4. We have by Claims 4.12 and 4.11 that, for at least $1-o(1)$ of uniformly random d_0-dimensional linear subspaces E from $(A_0)^\dagger$,

1. the fraction of $(d-d_0)$-dimensional linear subspaces $B' \supset E$ (from $(A_0)^\dagger$) that fall into $Cons$ is at least $\epsilon/6$, and
2. for every $x \in A_0 + E$, $\mathbf{Pr}_{B' \in Cons_E}[C(A_0+B')|_x = g(x)] \geqslant 1/10$.

Now consider the following distribution of subspaces E: pick a random $(d-d_0)$-dimensional linear subspace B satisfying the event of Eq. (10), and then pick a random d_0-dimensional linear subspace E inside B. By Lemmas 2.3 and 2.4, we conclude that when E is sampled according to this distribution, we get with probability at least $1/3-o(1)$ a subspace E such that both conditions (1) and (2) above still hold (assuming that $\epsilon > 1/k^{\Omega(1)}$ is large enough).

For subspaces B and $E \subset B$, we denote by $E' \subseteq A_0+E$ the subset of those $x \in A_0+E$ where $C(A_0+B)|_x \neq g(x)$. For every B satisfying the event of Eq. (10), we get by Lemma 2.2 that almost all subspaces $E \subset B$ are such that $|E'| > (\beta/2)|A_0+E|$. Combining this with our earlier argument, we get that for a random subspace B satisfying the event of Eq. (10), if we pick a random subspace $E \subset B$, we get with probability at least $1/3 - o(1) > 1/4$, a subspace E such that conditions (1) and (2) above hold, and additionally, $|E'| > (\beta/2)|A_0+E|$.

Fix any subspace E that satisfies the three conditions stated above. Let $E' \subset A_0+E$ be as above. By condition (2), we have that, for every $x \in E'$, there are at least $1/10$ fraction of subspaces $B' \in Cons_E$ such that $C(A_0+B')|_x = g(x) \neq C(A_0+B)|_x$. By averaging, for at least $1/20$ fraction of $B' \in Cons_E$, we have $C(A_0+B')|_x \neq C(A_0+B)|_x$ for at least $1/20$ fraction of $x \in E'$. Since we also know that $|E'| > |A_0+E|\beta/2$, we get that

$$C(A_0+B')|_{A_0+E} \overset{>\beta/40}{\neq} C(A_0+B)|_{A_0+E},$$

for at least $1/20$ fraction of $B' \in Cons_E$. By condition (1) on our fixed subspace E, we have that $Cons_E$ has measure at least $\epsilon/6$, and so

$$\mathbf{Pr}_{B':E\subset B'}[B' \in Cons \;\&\; C(A_0+B')|_{A_0+E} \overset{>\beta/40}{\neq} C(A_0+B)|_{A_0+E}] \geqslant \epsilon/120. \tag{11}$$

Since, for a random B conditioned on satisfying the event of Eq. (10), there are at least $1/4$ fraction of subspaces E such that Eq. (11) holds, we obtain

$$\mathbf{Pr}_{B,E\subset B,B'\supset E}[B' \in Cons \;\&\; C(A_0+B')|_{A_0+E} \overset{>\beta/40}{\neq} C(A_0+B)|_{A_0+E} \;\mid\; B \in Cons \;\&\; C(A_0+B) \overset{>\beta}{\neq} g(A_0+B)] \geqslant \epsilon/480,$$

where the probability is over picking a random $(d-d_0)$-dimensional linear subspace B (from $(A_0')^\dagger$) first, then picking its random d_0-dimensional linear subspace E, and finally picking a random

$(d-d_0)$-dimensional linear subspace B' (from $(A_0')^\dagger$) that contains E. Lifting the conditioning on B, we get

$$\mathbf{Pr}_{E,B\supset E,B'\supset E}[B' \in Cons \;\&\; C(A_0{+}B')|_{A_0+E} \overset{>\beta/40}{\neq} C(A_0{+}B)|_{A_0+E} \;\&\; B \in Cons] \geqslant \nu'\epsilon/480 > \nu\epsilon^2/960,$$

which contradicts the (α,γ)-excellence property for $\alpha = \beta/40$ and $\gamma = \nu\epsilon^2/960$. For $\gamma < \epsilon^3/960$, we get that $\nu < \epsilon$, as required. □

4.3 Local agreement implies global agreement

Here we conclude the analysis of the derandomized Z-Test by proving the following.

Lemma 4.13. *If the derandomized Z-test accepts with probability at least ϵ, then there is a function $g : \mathcal{U} \to \mathcal{R}$ such that for at least $\epsilon' = \epsilon/4$ fraction of all subspaces S, the oracle $C(S)$ agrees with $g(S)$ in all but at most $\alpha' = 81\alpha$ fraction of points $x \in S$.*

Proof. Let (A_0, B_0) be randomly chosen in the first step of the test. Let κ be the measure of the set $Cons_{A_0,B_0}$. If this (A_0, B_0) is not good (i.e., if $\kappa < \epsilon/2$), then the set B_1 chosen in the second step of the test will be in $Cons_{A_0,B_0}$ with probability at most $\epsilon/2$. If this happens, then the test may accept, but only with the probability less than $\epsilon/2$.

Thus we need to analyze the case where (A_0, B_0) is a random good pair, and so $\kappa \geqslant \epsilon/2$. By Corollary 4.3, all but at most ϵ^2 of good pairs (A_0, B_0) are excellent. If our chosen good pair(A_0, B_0) is not excellent, then the test may accept, but this happens with probability at most $\epsilon^2 < \epsilon$.

We are left with the case where our chosen pair (A_0, B_0) is (α,γ)-excellent. For this pair, we define the function g as the majority function over subspaces in $Cons_{A_0,B_0}$. By Lemma 4.4, we know that C mostly agrees with the direct product of g on almost all subspaces $(A_0 + B)$, where $B \in Cons_{A_0,B_0}$. We will argue that C will mostly agree with g^k also globally, on at least ϵ' fraction of all k-size subspaces S.

Consider picking subspaces B_1 and A_1 as follows: Pick a random d-dimensional subspace S, then randomly choose a $(d-d_0)$-dimensional subspace $B_1 \subset S$, and set A_1 to be any d_0-dimensional subspace of S that is linearly independent of B_1. This choice of B_1 and A_1 is essentially equivalent to the way they are chosen by the test. The only difference is that the subspace B_1 chosen by the test is linearly independent from the subspace A_0, whereas in our new way of picking B_1 it may happen that B_1 intersects A_0 in some non-zero point. However, since B_1 is uniformly distributed, the probability that it intersects A_0 in a non-zero point is negligible (less than $q^d/|\mathcal{U}|$). We will ignore this negligible amount, and think of this new choice of B_1 and A_1 as actually equivalent to the choices in the test.

For the sake of contradiction, suppose that there are fewer than ϵ' subspaces S where C and g^k have agreement in more than $1-\alpha'$ fraction of positions. Consider picking a random d-dimensional subspace S. If S is one of these ϵ' sets, then Test may accept, but this happens only with probability $\epsilon' < \epsilon$. So we may assume that S is a random subspace that contains more than α' fraction of inputs x where $C(S)|_x \neq g(x)$. Note that the distribution of such subspaces S is at most ϵ' far in statistical distance from the uniform distribution over all d-dimensional subspaces S.

For a random S, pick its random subspaces B_1 and A_1 as above. If $B_1 \notin Cons_{A_0,B_0}$, Test will reject. So the probability that Test accepts is at most the probability that Test accepts conditioned on $B_1 \in Cons_{A_0,B_0}$, times the probability that $B_1 \in Cons_{A_0,B_0}$.

By Lemma 4.4, all but at most $\nu = O(\gamma/\epsilon^2)$ fraction of subspaces $B \in Cons_{A_0,B_0}$ are such that $g(B)$ agrees with $C(A_0, B)|_B$ in all but at most $\beta = 40\alpha$ fraction of inputs $x \in B$. Conditioned

on choosing a $B_1 \in Cons_{A_0,B_0}$, the test chooses one of these ν fraction of sets with probability at most $(\nu\kappa + \epsilon')/(\kappa - \epsilon')$. Indeed, for a uniformly random subspace S, a random subspace $B_1 \subset S$ is uniformly distributed, and so hits the ν-fraction of $Cons_{A_0,B_0}$ with probability at most $\nu\kappa$. For the distribution of subspaces S that is ϵ'-far from uniform, this hitting probability may increase by at most ϵ'. On the other hand, the probability that $B_1 \in Cons_{A_0,B_0}$ is at least $\kappa - \epsilon'$. The claimed bound follows.

At the same time, since $C(S) \overset{>\alpha'}{\neq} g(S)$, we get by Lemma 2.2 that, for all but at most $O(q^2/(|B_1|\alpha'))$ of subspaces B_1 of S,

$$C(S)|_{B_1} \overset{>\alpha'/2}{\neq} g(B_1). \tag{12}$$

Conditioning on $B_1 \in Cons_{A_0,B_0}$ for a random S means that this probability gets multiplied by at most $1/(\kappa - \epsilon')$.

So conditioned on $B_1 \in Cons_{A_0,B_0}$, the probability that $B_1 \subset S$ for a random S was chosen so that either (12) is violated or that $g(B_1) \overset{>\beta}{\neq} C(A_0, B_1)|_{B_1}$ is at most

$$\rho = (\nu\kappa + \epsilon' + O(q^2/(k\alpha)))/(\kappa - \epsilon').$$

In this case, the test may accept, but only with probability at most $\rho \cdot \mathbf{Pr}[B_1 \in Cons_{A_0,B_0}] \leqslant \rho(\kappa + \epsilon')$. This probability is less than ϵ, since $(\kappa + \epsilon')/(\kappa - \epsilon') \leqslant 3$ for $\epsilon' = \epsilon/4$ and $\kappa \geqslant \epsilon/2$.

Finally, for B_1 that satisfies (12) and is such that $g(B_1) \overset{\geqslant 1-\beta}{=} C(A_0, B_1)|_{B_1}$, we get that

$$C(A_1, B_1)|_{B_1} \overset{>\alpha'/2-\beta}{\neq} C(A_0, B_1)|_{B_1}.$$

Since $\alpha'/2 > \beta$, the test will reject in this case.

Thus we have argued that, in all cases, the probability that the test accepts is strictly less than ϵ. Hence, the function g must be an approximate direct-product function for C on ϵ' fraction of all d-dimensional subspaces. □

Proof of Theorem 1.2. The proof easily follows from Lemma 4.13 above. □

4.4 Two queries suffice for the derandomized case

Using the same arguments as in Section 3.4, we also prove Theorem 1.3. The proof is along the lines of the proof of Theorem 3.17. The only change is the use of Chebyshev's instead of Chernoff-Hoeffding's inequalities (using pairwise independence of random linear subspaces).

5 DP testing for non-Boolean functions

In this section, we generalize Lemma 3.10 to the non-Boolean case. The proof is a reduction to the Boolean case that gets a slightly weaker value of β.

Lemma 5.1. *Let $\mathcal{R}$ be an arbitrary finite set. If $C' : Cons \to \mathcal{R}^k$ is (α, γ)-excellent with respect to $Cons$, and G is its plurality function, then there are fewer than $\nu = O(\gamma/\epsilon^2) < \epsilon$ fraction of sets $B \in Cons$ such that $C'(B)|_x \neq G(x)$ for more than $\beta = 320\alpha$ fraction of $x \in B$.*

Proof. Let Bad be the collection of those sets $B \in Cons$ such that $G(B)$ and $C(A_0, B)|_B$ disagree in at least 320α fraction of positions.

Let $F = \{F_x\}_{x \in \mathcal{U}}$ be a family of independent random functions $F_x : \mathcal{R} \to \{0,1\}$. Define $C_F(x_1, \ldots, x_k) = F_{x_1}(y_1), \ldots, F_{x_k}(y_k)$, where $y_1, \ldots, y_k = C'(x_1, \ldots, x_k)$. In other words, C_F takes the values $y_1, \ldots, y_k$ returned by C' on $x_1, \ldots, x_k$, and maps them into the Boolean values by applying F_{x_i} to y_i, for $1 \leqslant i \leqslant k$.

Observe that if C' is consistent on two overlapping sets, so will be new function C_F on the same sets. So, for each fixed family F, we get that C_F is excellent with respect to the $Cons$. Hence, by Lemma 3.10, we have that for almost all $B \in Cons$, $C'(B)$ is close to the Boolean majority function g_F on B. In particular, the set Bad_F (of sets in $Cons$ that disagree in 40α fraction of positions with g_F) has measure less than ν as above.

On the other hand, we will show that the expected size of Bad_F is at least almost the same as that of Bad. (Note that, since G is not a majority, but just a plurality, we needn't have $F_x(G(x)) = g_F(x)$, so Bad_F may not be contained in Bad.) Indeed, fix an x, and consider an arbitrary u such that $u \neq G(x)$. Define a random function F_x on all elements in $\mathcal{R}$, except for u and $G(x)$. Let $b \in \{0,1\}$ be the majority value $g_F(x)$ so far (i.e., the majority of the values of $F_x(r)$, for $r = C'(A_0, B)|x$ for $B \in Cons$ containing x, where $r \in \mathcal{R}$ is not u or $G(x)$). Since F_x independently maps u and $G(x)$ to $\{0,1\}$, we get with probability $1/4$ that $G(x)$ is mapped to b and u is mapped to $1-b$. But $G(x)$ is more popular than u, and so b will be equal to $g_F(x)$. Hence, the probability (over the choice of F) that $g_F(x) = F_x(u)$ is at most $1/4$.

Recall that every set $B \in Bad$ has 320α fraction of elements x where $C'(B)|_x \neq G(x)$. By the above, each such element has a $1/4$ chance of having $C_F(B)_x \neq g_F(x)$. Thus, almost surely (with probability at least $1 - e^{-\Omega(\alpha k)}$), $B \in Bad_F$. Therefore, the expected size of Bad_F is at least almost the same as that of Bad. But for each F, Bad_F has measure less than ν (by the Boolean case analysis of Lemma 3.10). Therefore, Bad has measure less than $O(\nu)$, as desired. □

Using this new lemma in place of Lemma 3.8 for the case of non-Boolean functions with an arbitrary range $\mathcal{R}$, we get that our DP testing results (Theorems 1.1 and 1.2) continue to hold in this case.

6 A 2-query PCP, and a new parallel repetition theorem

Here we analyze the 2-query PCP construction given in the Introduction (Theorem 1.4). We then show (in Section 6.2) how our analysis can be viewed as a parallel repetition theorem for certain 2-prover games.

6.1 Proof of Theorem 1.4

Here we give a generic reduction from a graph CSP (G, Φ) over an alphabet Σ, with completeness σ and soundness $1-\delta$, to a 2-query PCP over the alphabet Σ^k with completeness $1-\exp(-\sigma k)$ and soundness $1 - \exp(-\delta k')$, for $k' = \Theta(\sqrt{k})$. Throughout this section, we identify U (the vertex set of the CSP graph G) with the universe $\mathcal{U}$, and the alphabet Σ with the range $\mathcal{R}$ (to be consistent with the notation used earlier in the paper for direct product testing).

We first re-state the description of our verifier $\mathcal{Y}$ and Theorem 1.4 from the Introduction. Recall that we define the PCP proof to be a function C_E that, given a set of k edges in the constraint graph G, returns assignments to all of the end-points of these edges. Let $k' < k$ be the parameter from our DP test above. Our 2-query verifier is the following.

Verifier $\mathcal{Y}$:

1. Pick a set of k' random vertices A. For each vertex $v \in A$, pick a random incident edge (v, v') in G. Let $A_{E,1}$ be the set of these k' edges. Independently, pick another set $A_{E,2}$ of k' random edges incident on the vertices in A. Finally, pick two random sets of edges $B_{E,1}$ and $B_{E,2}$, of size $k - k'$ each.
2. Query $C_E(A_{E,1}, B_{E,1})$ and $C_E(A_{E,2}, B_{E,2})$. Accept iff the following checks pass:
(a) the query answers satisfy $0.9 \cdot \sigma$ fraction of constraints on each of the $B_{E,i}$'s, and
(b) they assign the same values to A.

Theorem 6.1. *(i) If a CSP-instance (G, Φ) is σ-satisfiable, then there is a proof C_E accepted by verifier $\mathcal{Y}$ with probability $\sigma' \geqslant 1 - \exp(-\sigma k)$; moreover, if $\sigma = 1$, then $\sigma' = 1$. (ii) If the CSP-instance is δ-unsatisfiable, then no proof C_E is accepted by $\mathcal{Y}$ with probability greater than $\epsilon = e^{-(1/c)\delta k'}$, for some fixed constant c.*

Proof. For part *(i)*, an honest proof C_E (based on some σ-satisfying assignment for (G, Φ)) will be accepted with the stated probability σ', by the Chernoff bounds. Moreover, if $\sigma = 1$, then the honest proof will be accepted with probability $\sigma' = 1$.

For part *(ii)*, intuitively, we will argue that the consistency of the proof C_E on a vertex set A implies the existence of an assignment $g : U \to \Sigma$ consistent with C_E. But no assignment can satisfy significantly more than δ fraction of the random edge constraints of $B_{E,2}$ (by the soundness assumption). Therefore C_E will be rejected by $\mathcal{Y}$. We provide the details next.

Let us define (for the sake of the analysis only) a probabilistic function C from k sets of vertices to $\mathcal{R}^k$ as follows: Given a k-size vertex-set S, pick k edges S_E at random, one incident to each node in S. Output $C_E(S_E)|_S$.

Imagine applying our DP testing analysis (from Section 3) to this function C. The V-test with respect to C is as follows: Pick a random k'-size vertex-set A, pick random $(k - k')$-size vertex sets B_1 and B_2 at random, and then check whether $C(A, B_1)|_A = C(A, B_2)|_A$. Note that this is exactly the same as the consistency check done in Step 2(b) of our verifier $\mathcal{Y}$ above. (Indeed, C would pick random edges $A_{E,1}$ and $A_{E,2}$ incident to A, and then random edges incident to each of B_i, $i = 1, 2$. The latter are just sets of random edges, since the graph is regular, and so have the same distribution as $B_{E,i}$.)

Let a be the values assigned to A by $C_E(A_{E,1}, B_{E,1})$ in Step 2 of verifier $\mathcal{Y}$. For δ and ϵ in the statement of the present theorem, we set $\alpha = \delta/320$ and $\gamma = \epsilon^4$. We classify pairs (A, a) as being good, (α, γ)-excellent, or neither, with respect to C, using the corresponding definitions from Section 3[13].

We consider three ways that verifier $\mathcal{Y}$ may accept the given proof C_E:

1. (A, a) is not good. Then the conditional probability of passing the consistency check in Step 2(b) is the probability that $C_E(A_{E,2}, B_{E,2})|_A = a$. This is the same as the probability that $C(A, B_2)|_A = a$, which is at most $\epsilon/2$ by the definition of goodness.

2. (A, a) is good but not excellent. By Lemma 3.6, the probability that (A, a) is good but not (α, γ)-excellent is less than $e^{-\Omega(\alpha k')}/\gamma$, which can be made less than $\epsilon/4$ by choosing a sufficiently large constant c (in the statement of the present theorem); here and below we also use our assumption that $\epsilon < 1/4$.

[13] with a natural modification to allow randomized oracles C; so all the probabilities are now also over the internal randomness of the oracle C being tested.

3. (A, a) *is excellent.* By Lemma 3.8, there is a function $g = g_{A,a} : U \to \Sigma$,[14] so that

$$\mathbf{Pr}_B[C(A,B)|_A = a \;\&\; C(A,B)|_B \overset{>40\alpha}{\neq} g(B)] < \gamma/\epsilon^2 = \epsilon^2,$$

where the probability is over random $(k-k')$-size vertex sets $B \subseteq U \setminus A$, and internal randomness of C. Making the internal randomness of C explicit, we can re-write the probability above as follows:

$$\mathbf{Pr}_{A_{E,2},B_{E,2},B}[C_E(A_{E,2},B_{E,2})|_A = a \;\&\; C_E(A_{E,2},B_{E,2})|_B \overset{>40\alpha}{\neq} g(B)] < \epsilon^2, \tag{13}$$

where $A_{E,2}$ is the set of random edges incident on A, the set $B_{E,2}$ is the set of $(k-k')$ random edges (as chosen by our verifier $\mathcal{Y}$), and B is the set of vertices obtained by randomly selecting an end-point from every edge in $B_{E,2}$. (Note, thanks to the regularity of the graph G, this way of choosing $B_{E,2}, B$ is the same as choosing a k'-size vertex set B first and then choosing its random incident edges $B_{E,2}$.)

We claim that

$$\mathbf{Pr}_{A_{E,2},B_{E,2}}[C_E(A_{E,2},B_{E,2})|_A = a \;\&\; C_E(A_{E,2},B_{E,2})|_{B_{E,2}} \overset{>100\alpha}{\neq} g(B_{E,2})] < \epsilon^2 + \exp(-\alpha k).$$

Indeed, suppose otherwise. Condition on any $A_{E,2}, B_{E,2}$ satisfying the random event in the above probability expression. Pick B by randomly selecting an end-point from every edge in $B_{E,2}$. Every edge in $B_{E,2}$ where C_E and g disagree will contribute, with probability at least $1/2$, a vertex to B where C_E and g disagree. (This is because at least one of the end-points of this edge is in disagreement with g.) By Chernoff, the probability that B contains fewer than 40α fraction of vertices where C_E and g disagree is less than $\exp(-\alpha k)$. But then we get a contradiction to Eq. 13 above.

Finally, by the soundness assumption for (G, Φ), every assignment violates at least δ fraction of edge constraints in G. In particular, this is true for our function g. The $(k-k')$ edges in $B_{E,2}$ are random and independent edges in G. By Chernoff, the probability that fewer than $\delta/2$ fraction of them have their constraints violated by g is $e^{-\Omega(\delta\cdot(k-k'))} < \epsilon/8$.

Assuming that none of the low-probability events above happened, we get that the answers $C_E(A_{E,2}, B_{E,2})$ violate at least $\delta/2 - 100\alpha = (3/16)\delta$ fraction of the edges in $B_{E,2}$. But then verifier $\mathcal{Y}$ would reject. It follows that the verifier may accept with probability at most $\epsilon/2 + \epsilon/4 + \epsilon^2 + \exp(-\alpha k) + \epsilon/8 < \epsilon$, as required. □

Remark 6.2. The value k' must satisfy the condition that $k'^2 \leqslant O(k)$. So we can choose $k' = \Theta(\sqrt{k})$. Then the ϵ in the statement of Theorem 6.1 becomes $e^{-\Omega(\delta\sqrt{k})}$.

6.2 A new parallel repetition theorem

Our 2-query PCP from the previous subsection may be viewed as a new parallel repetition theorem for a certain family of 2-prover games. These restricted games are closely related to general games and we will explain the connection later.

Let $G(V, E)$ be a d-regular graph, and let $C : E \to 2^{\Sigma^2}$ be a set of edge constraints.

The usual constraint satisfaction problem $S = S(G, C)$ asks for a labeling of the vertices by symbols from Σ which maximizes the number of edges whose constraints are satisfied. The game

[14] Here, for $x \in U \setminus A$, $g(x)$ is defined to be the most likely value $C(A,B)|_x$, over random $(k-k')$-size vertex-sets B containing x (and internal randomness of C), conditioned on $C(A,B)|_A = a$; if no such value exists for x, we set $g(x)$ to equal some default symbol in Σ.

S may be viewed as a 2-prover game, in which a verifier picks an edge from E at random, gives an endpoint to each player and verifies that the answer (the labels proposed by the provers) satisfies that edge constraint. The value of the game S, i.e., the maximum probability of the provers to satisfy the verifier, is essentially the maximum fraction of edges satisfied by the optimal assignment.

But one can define another game, $T = T(G, C)$ with similar connection to the given CSP. Here the verifier picks a *pair* of edges at random (from some distribution P), sends one edge each of the provers, and checks two things about the answers (that label the endpoints of each edge): *(a)* the edge constraints are satisfied, and *(b)* if the two edges share a vertex, the labels given to that vertex are the same.

The most natural (and used) distribution P is to pick a pair of incident edges uniformly at random (so condition *(b)* always applies). In this case the value of the game $T[P]$ is essentially the same as that of the game S.

Here is another natural distribution Q: pick the two edges uniformly at random. In this case, condition *(b)* almost never applies, and the value of the game $T[Q]$ is almost 1.

The family of games we will consider use a mixture of these two distributions, $pP + qQ$ with $p + q = 1$. In particular, we use $p = 1/m$. Note that if the value of the game with P is $1 - v$, then the value of the new game $T[(P + (m-1)Q)/m]$ is $1 - (v/m)$.

While "diluting" the quality of the game, the advantage of the mixture is in making it hard for the players to coordinate. In particular, the famous counterexamples of Feige and Verbitsky[FV02] and of Raz [Raz08] don't seem to hold for such games. Indeed, we get the following.

Theorem 6.3. *For $k = m^2$, the value of the game $T[(P + (m-1)Q)/m]^k$ (the game T repeated k times, in the standard sense of parallel repetition) is $(1 - (v/m))^{\Omega(k)}$.*

Proof. It follows immediately from the proof of Theorem 6.1, as the k-tuple of pairs of questions in the game $T[(P + (m-1)Q)/m]$ will almost certainly yield about $\sqrt{k}$ pairs of incident edges, with the rest being pairs of independent edges. Hence, the analysis of the verifier $\mathcal{Y}$ applies. □

Note that the value of the k-wise parallel repetition of the game $T[(P+(m-1)Q)/m]$ is roughly $(1 - v)^m$ (ignoring a constant in the exponent), and this bound does not depend on Σ! We hope that we can take m to be an absolute constant, independent of k, and still have perfect decay $(1 - v)^k$.

6.3 The Feige-Kilian parallel repetition: Proof of Theorem 1.5

First, we recall the definition of our verifier $\mathcal{Y}'$ and re-state Theorem 1.5. Let (G, Φ) be a graph CSP with the vertex set U and the alphabet Σ. The first prover C_1 gets as input a k'-subset of vertices of G and returns an assignment to all these vertices. The second prover is a function C_E that, given a set of k edges of G, returns assignments to all the $2k$ end-points of these edges.

Verifier $\mathcal{Y}'$:
1. Pick a set of k' random vertices A. For each vertex $v \in A$, pick a random incident edge (v, v') in G. Let $A_{E,2}$ be the set of these k' edges. Pick a set of $(k - k')$ random edges $B_{E,2}$.
2. Query $C_1(A)$ and $C_E(A_{E,2}, B_{E,2})$. Accept iff the following checks pass:
(a) the query answers satisfy $0.9 \cdot \sigma$ fraction of constraints of $B_{E,2}$, and
(b) they assign the same values to A.

Theorem 6.4. *(i) If a CSP-instance (G, Φ) is σ-satisfiable, then there are proofs (C_1, C_E) accepted by verifier $\mathcal{Y}'$ with probability $\sigma' \geqslant 1 - \exp(-\sigma k)$; moreover, if $\sigma = 1$, then $\sigma' = 1$. (ii) If the CSP-instance is δ-unsatisfiable, then no proofs (C_1, C_E) are accepted by $\mathcal{Y}'$ with probability greater than $\epsilon = e^{-(1/c)\delta k'}$, for some fixed constant c.*

Proof sketch. First we observe that our analysis of the V-test can be easily adapted to the scenario where the two queries are made to two different provers. The first prover C_1 gives an assignment for k'-subsets of the universe $\mathcal{U}$, and the second prover C_2 gives an assignment for k-subsets of $\mathcal{U}$. The test picks a random k'-subset $A_0 \subseteq \mathcal{U}$ and a random k-subset $(A_0, B_1) \subseteq \mathcal{U}$, and accepts if $C_1(A_0) = C_2(A_0, B_1)|_{A_0}$.

In this new setting, we define the set $Cons_{A_0}$ as the set of all those $k - k'$-subsets B where $C_1(A_0) = C_2(A_0, B)|_{A_0}$. We call a set A_0 *good* if the measure of $Cons_{A_0}$ is at least $\epsilon/2$. We call A_0 (α, γ)-*excellent* if it is good and

$$\mathbf{Pr}_{E,D_1,D_2}[(E, D_i) \in Cons_{A_0},\ i = 1, 2,\ \&\ C_2(A_0, E, D_1)|_E \overset{>\alpha}{\neq} C_2(A_0, E, D_2)|_E] \leqslant \gamma.$$

One can easily check that all lemmas in Sections 3.1 and 3.2 continue to hold for this new test (with the same proofs). That is, we get the following: *(1)* if the new test accepts with probability at least ϵ, then a random subset A_0 is good with probability at least $\epsilon/2$; *(2)* the probability that A_0 is good but not (α, γ)-excellent is less than γ'/γ, where $\gamma' = \exp(\alpha k')$; and *(3)* for any excellent A_0 and the corresponding plurality function $g = g_{A_0}$ (defined with respect to $Cons_{A_0}$), there are fewer than $\nu = O(\gamma/\epsilon^2)$ fraction of sets $B \in Cons_{A_0}$ such that $C_2(A_0, B)|_x \neq g(x)$ for more than 40α fraction of $x \in B$, where $\alpha > \Omega((\ln 1/\epsilon)/(k/k'))$.

Now the analysis of the verifier $\mathcal{Y}'$ is very similar to that of the verifier $\mathcal{Y}$ given in Section 6.1. We just define the randomized "vertex-proof" C_2 from k-sets of vertices to Σ^k as follows: Given a k-size vertex set S, pick at random k edges S_E, one incident edge per node in S; output $C_E(S_E)|_S$. Then we observe that the test $\mathcal{Y}'$ is applying (the 2-prover version of) the V-test to the provers C_1 and C_2. The rest of the argument is exactly the same as in Section 6.1. □

7 Open questions

While our current techniques stop at the exponent $\sqrt{k}$ in the soundness decay, we see no obvious obstacle to improving it to k, and proving possibility/impossibility of this is one interesting open question we leave. Another interesting question is whether our PCP construction works for $k < 1/\delta^2$; our current analysis seems to require that $k > 1/\delta^2$. Perhaps the most interesting open question is whether our techniques can be used to construct a 2-query PCP with *sub-constant* soundness, thereby providing an alternative construction to [MR08].

Acknowledgments We wish to thank Irit Dinur, Ariel Gabizon, Oded Goldreich, Or Meir, Dana Moshkovitz, Ryan O'Donnell, and Anup Rao for their comments on an early version of this paper. We thank Sanjeev Arora for making us realize that our 2-query PCP verifier $\mathcal{Y}$ is essentially the same as the miss-match proof system of Feige and Kilian [FK00], which led us to the proof of Theorem 1.5.

References

[ALM+98] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *Journal of the Association for Computing Machinery*, 45(3):501–555, 1998.

[AS98] S. Arora and S. Safra. Probabilistic checking of proofs: A new characterization of NP. *Journal of the Association for Computing Machinery*, 45(1):70–122, 1998.

[DG08] I. Dinur and E. Goldenberg. Locally testing direct products in the low error range. In *Proceedings of the Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science*, pages 613–622, 2008.

[Din07] I. Dinur. The PCP theorem by gap amplification. *Journal of the Association for Computing Machinery*, 54(3), 2007.

[Din08] I. Dinur. PCPs with small soundness error. *ACM SIGACT News*, 2008. Guest complexity theory column.

[DR06] I. Dinur and O. Reingold. Assignment testers: Towards a combinatorial proof of the PCP theorem. *SIAM Journal on Computing*, 36(4):975–1024, 2006.

[EKK+00] F. Ergün, S. Kannan, R. Kumar, R. Rubinfeld, and M. Viswanathan. Spot-checkers. *Journal of Computer and System Sciences*, 60(3):717–751, 2000.

[FK00] U. Feige and J. Kilian. Two-prover protocols - low error at affordable rates. *SIAM Journal on Computing*, 30(1):324–346, 2000.

[FV02] U. Feige and O. Verbitsky. Error reduction by parallel repetition - A negative result. *Combinatorica*, 22(4):461–478, 2002.

[GL89] O. Goldreich and L.A. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, pages 25–32, 1989.

[GNW95] O. Goldreich, N. Nisan, and A. Wigderson. On Yao's XOR-Lemma. *Electronic Colloquium on Computational Complexity*, TR95-050, 1995.

[GS00] O. Goldreich and S. Safra. A combinatorial consistency lemma with application to proving the PCP theorem. *SIAM Journal on Computing*, 29(4):1132–1154, 2000.

[Hol07] T. Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing*, pages 411–419, 2007.

[IJK06] R. Impagliazzo, R. Jaiswal, and V. Kabanets. Approximately list-decoding direct product codes and uniform hardness amplification. In *Proceedings of the Forty-Seventh Annual IEEE Symposium on Foundations of Computer Science*, pages 187–196, 2006.

[IJKW08] R. Impagliazzo, R. Jaiswal, V. Kabanets, and A. Wigderson. Uniform direct-product theorems: Simplified, optimized, and derandomized. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 579–588, 2008. (full version available as *ECCC TR08-079*).

[Imp02] R. Impagliazzo. Hardness as randomness: A survey of universal derandomization. *Proceedings of the ICM*, 3:659–672, 2002. (available online at arxiv.org/abs/cs.CC/0304040).

[IW97] R. Impagliazzo and A. Wigderson. P=BPP if E requires exponential circuits: Derandomizing the XOR Lemma. In *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, pages 220–229, 1997.

[Lev87] L.A. Levin. One-way functions and pseudorandom generators. *Combinatorica*, 7(4):357–363, 1987.

[MR08] D. Moshkovitz and R. Raz. Two query PCP with sub-constant error. In *Proceedings of the Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science*, 2008.

[Rao08] A. Rao. Parallel repetition in projection games and a concentration bound. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, pages 1–10, 2008.

[Raz98] R. Raz. A parallel repetition theorem. *SIAM Journal on Computing*, 27(3):763–803, 1998.

[Raz08] R. Raz. A counterexample to strong parallel repetition. In *Proceedings of the Forty-Ninth Annual IEEE Symposium on Foundations of Computer Science*, 2008.

[Tre03] L. Trevisan. List-decoding using the XOR lemma. In *Proceedings of the Forty-Fourth Annual IEEE Symposium on Foundations of Computer Science*, pages 126–135, 2003.

[Ver96] O. Verbitsky. Towards the parallel repetition conjecture. *Theoretical Computer Science*, 157(2):277–282, 1996.

[Yao82] A.C. Yao. Theory and applications of trapdoor functions. In *Proceedings of the Twenty-Third Annual IEEE Symposium on Foundations of Computer Science*, pages 80–91, 1982.

A Analysis of the Z'-test: Proof of Theorem 3.21

We'll show first that the assumption of Theorem 3.21 implies that with almost the same probability, both (A_0, B_0) and (A_1, B_1) are excellent, and furthermore, the respective plurality functions G_0 and G_1 are close to each other. Then we argue that for some fixed (A_0, B_0), for the plurality function G_0, G_0^k is close to C on a poly(ϵ') fraction of assignments. A more formal proof is given next.

For the rest of the proof, assume that $p = 2\epsilon'$. Let $\alpha = C(\log 1/\epsilon')k/k'$ for a suitably large constant C. Let $\beta = 320\alpha$. Let $\gamma = 4 \cdot e^{-c\alpha k'}/\epsilon'$ for c the hidden constant in the expression for γ' from Lemma 3.6, so that $\gamma'/\gamma < \epsilon/4$. By picking C suitably large, we can assume $\gamma = o(\epsilon'^3)$.

Claim A.1. *The probability that C passes the Z'-test and both (A_0, B_0) and (A_1, B_1) are good is at least ϵ'.*

Proof. Since $N \cup A_1$ is independent of A_0, and passing the Z'-test implies that $N \cup A_1 \in Cons_{A_0,B_0}$, the probability of passing given that (A_0, B_0) is not good is at most $\epsilon'/2$, as is the probability of (A_0, B_0) not being good and C passing the Z'-test. Similarly for (A_1, B_1) not being good and C passing the Z'-test. Subtracting the probability of both of these events from the $2\epsilon'$ probability of passing the Z' test gives the claim. □

Claim A.2. *The probability that C passes the Z'-test and both (A_0, B_0) and (A_1, B_1) are (α, γ)-excellent is at least $\epsilon'/2$.*

Proof. By Lemma 3.6, the probability that (A_0, B_0) is good but not (α, γ)-excellent is at most $\gamma'/\gamma = \epsilon'/4$ by our choice of parameters, and similarly for (A_1, B_1). Subtracting these two bad events from the probability of Claim A.1 yields the bound. □

Claim A.3. *The probability over sets A_0, B_0, A_1, B_1, N that both (A_0, B_0) and (A_1, B_1) are excellent, and, for the respective plurality functions G_0, G_1,*

$$G_0^{k-2k'}(N) \overset{\geq 1-2\beta-2k'/k}{=} G_1^{k-2k'}(N)$$

is at least $\epsilon'/4$.

Proof. Consider the event that (A_0, B_0) is excellent, $N \cup A_1 \in Cons_{A_0,B_0}$ and $G_0(N \cup A_1) \overset{\geq \beta}{\neq} C(A_0 \cup N \cup A_1)_{N \cup A_1}$. By Lemma 5.1, this event occurs with probability at most $\nu = O(\gamma/\epsilon'^2) = o(\epsilon')$. Similarly for the corresponding event concerning (A_1, B_1) and $N \cup A_0$ for the function G_1. If neither of these two events occurs, then $G_0^{k-2k'}(N) \overset{\geq 1-\beta-k'/k}{=} C(A_0 \cup A_1 \cup N)_N \overset{\geq 1-\beta-k'/k}{=} G_1^{k-2k'}(N)$, so $G_0^{k-2k'}(N) \overset{\geq 1-2\beta-2k'/k}{=} G_1^{k-2k'}(N)$. Thus, we get the bound in the claim by subtracting two $o(\epsilon')$ events from the $\epsilon'/2$ probability event in Claim A.2. □

Claim A.4. *The probability over sets A_0, B_0, A_1, B_1 that (A_0, B_0), (A_1, B_1) are both excellent, and $G_0(x) = G_1(x)$ for all but a 4β fraction of x's is at least $\epsilon'/8$.*

Proof. Consider the event that $G_0(x) \neq G_1(x)$ for a 4β fraction of x, but $G_0^{k-2k'}(N) \overset{\geq 1-2\beta-2k'/k}{=} G_1^{k-2k'}(N)$. For any fixed (A_0, B_0) and (A_1, B_1) with G_0, G_1 that distant, since N is uniformly distributed subset of U, we get by Chernoff bounds that the likelihood of almost equality on N is $e^{-O(\beta k)} = o(\epsilon')$ by the choice of parameters. Subtracting this probability from that in Claim A.3 gives the bound. □

Claim A.5. *There exists an excellent (A_0, B_0) so that with probability at least $\Omega(\epsilon'^2)$ over k-sets S , $G_0^k(S) \overset{\geq 1-O(\beta+k'/k)}{=} C(S)$.*

Proof. Pick any (A_0, B_0) so that the conditional probability of the event in the previous claim is at least $\epsilon'/8$. Pick S as follows: Pick random sets A_1, B_1 and B_2 of sizes k', $k - k'$ and $k - k'$, respectively. Let $S = A_1 \cup B_2$. Note that since $|A_1|$ is small, we only need to look at the circuit and G_0 on B_2. Then the probability that (A_1, B_1) is (α, γ)-excellent and G_1 is $O(\beta)$ close to G_0 is $\Omega(\epsilon')$. If (A_1, B_1) is excellent, and hence good, the conditional probability that $B_2 \in Cons_{A_1,B_1}$ is $\Omega(\epsilon')$. If this occurs, by Lemma 5.1, almost certainly $C(A_1, B_2)_{B_2} \overset{\geq 1-O(\beta+k'/k)}{=} G_1^{k-k'}(B_2)$. Thus, with probability $\Omega(\epsilon'^2)$, $C(S)_{B_2} \overset{\geq 1-O(\beta+k'/k)}{=} G_1^{k-k'}(B_2)$. Also, since G_1 and G_0 are close, with probability $1 - o(\epsilon'^2)$, $G_1^{k-k'}(B_2) \overset{\geq 1-O(\beta-k'/k)}{=} G_0^{k-k'}(B_2)$. The claim then follows. □

The proof of the theorem is immediate from the last claim.