New Direct-Product Testers and 2-Query PCPs

Russell Impagliazzo[*]
UC San Diego, La Jolla, CA &
Institute for Advanced Study
Princeton, NJ, USA
russell@cs.ucsd.edu

Valentine Kabanets[†]
Simon Fraser University
Burnaby, BC, Canada
kabanets@cs.sfu.ca

Avi Wigderson[‡]
Institute for Advanced Study
Princeton, NJ, USA
avi@ias.edu

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 [12] pioneered its local *testing* and its PCP application. A recent result by Dinur and Goldenberg [5] 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 [15]) also allow us to considerably simplify the analysis of the 2-query test of [5]. 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 [19] (building on Raz's (k-fold) parallel repetition theorem [20] and Holenstein's proof [13]) 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 of Feige and Kilian [8], but with simpler analysis and exponentially better soundness error.

Categories and Subject Descriptors

F.1.2 [**Theory of Computation**]: Modes of Computation; F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Nonnumerical Algorithms and Problems

General Terms

Algorithms, Theory

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 [20]. 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 [9] and the recent [21]) 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 [23]. A few years later Raz proved his parallel repetition theorem [20], showing that indeed the

[*] Work partially supported by NSF grants 0835373 and 0832797 and the Ellentuck Foundation

[†] Work partially supported by an NSERC Discovery grant

[‡] Work partially supported by NSF grants CCF-0832797 and DMS-0835373

STOC'09, May 31–June 2, 2009, Bethesda, Maryland, USA.

decay is exponential. Simpler proofs [13, 19] 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 [4]. The best current result is the tour-de-force of Moshkovitz and Raz [18]. Here we focus on using direct-product testing for this purpose, an idea pioneered by Goldreich and Safra [12]. 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 survey known 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 $C : \mathcal{U}^k \to \mathcal{R}^k$ we'd like 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 passes with a "significant" probability q, will certify that C is "sufficiently close" to $f^{(k)}$ for some 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 [7].

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 [12] described a test with a constant number of queries, and this was improved to the optimal two-query test by Dinur and Reingold [6]. 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. This is the "list-decoding" regime, which has been so important in recent developments in coding theory and complexity theory, and requires subtle tests and analysis even for far better codes than the direct-product code.

The first result to test the direct-product code in the list-decoding regime was obtained a few months ago by Dinur and Goldenberg [5] (building on the earlier work by Feige and Kilian [8]). They give a 2-query test which, if C passes with probability $q > 1/k^\alpha$ (for some fixed $\alpha > 0$), certifies that C can be approximately list-decoded as above. In particular, one codeword in the list must approximately agree with a q-fraction of the entries in C. 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 $q < 1/k$.

1.2 Our results

Our main result is that one *can* achieve exponential decay if allowed one additional query! We give a 3-query test which, if passed by C with probability $q > \exp(-k^{1/3})$, certifies that C can be approximately list-decoded, and in particular one codeword in the list approximately agrees with a q-fraction of the entries in C. Our techniques (see below) also allow us to considerably simplify the analysis of the 2-query test of [5] (for poly$(1/k)$-agreement).

To explain our next result, *derandomized* direct product testing, we revisit the PCP motivation, and another parameter of the amplification quality which we have ignored so far, namely proof size. 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 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, in that an oracle C which passes the test would still be close to an approximate direct product function.

Goldreich and Safra [12] 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 [5] as an open question. We solve it here for their 2-query test in the list-decoding regime! We show that, for *any* k, there is a polynomial family of k-tuples, such that if C passes the 2-query test of [5] 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, list-decodable k-fold direct-product code of polynomial rate.

Finally we return to the motivation of using direct-product testing to improve the soundness amplification of PCPs. The naive approach to do so would involve both the direct product test and then a parallel repetition test[3]. Here, we show,

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

[2] This in effect precludes inapproximability results to depend on $\mathsf{P} \neq \mathsf{NP}$

[3] Even this is not guaranteed to work directly from the definition of direct product testing. One needs that a strategy passing the direct product test is basically a probability dis-

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 [8]. 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 [8]. We also want to stress that our PCP analysis does not use our direct product testing result as a black-box. 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.

1.3 Our techniques, and DP decoding

The direct-product code we have been discussing all along has been under study in complexity theory long before PCPs were invented. Yao's XOR Lemma [24, 17], and its sibling, the "concatenation lemma" (proved equivalent by Goldreich and Levin [10]), served for almost three decades as the basic hardness amplification tool. It received many different proofs (e.g,. [11, 16]) worthy of its many uses.

Trevisan [22], motivated by proving the concatenation lemma in the "uniform setting", was the first to express it in the coding language as follows: Given C with the promise that it has q-agreement with some direct-product function, find a list of all such functions.

There is no clear reduction between direct product testing and direct product decoding.[4] 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) [14, 15] 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 [5] 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

1.4.1 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 $f : \mathcal{U} \to \mathcal{R}$.[5] 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.

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 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. The three intersecting sets (B_0, A_0), (A_0, B_1), and (B_1, A_1) can be pictured to form a shape of the letter "Z" – whence the name of the test.

If we restrict the test above to just the first two steps, we get the 2-query test analyzed by [6, 5]. We will call this 2-query test a *V-test* (as two intersecting sets can be pictured to form a letter "V"). As proved by [5], 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 [15]. 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[6] subspace A of $\mathcal{U}$, we set $f^k(A)$ to be the values of f on all $k = q^d$ points in the

tribution over independent strategies, rather than merely correlated with an independent strategy.

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

[5] Think of Boolean functions f for simplicity. However, as we show in the full version of the paper, our tests work for arbitrary ranges $\mathcal{R}$.

[6] [15] uses *affine* subspaces, but one could also use just linear subspaces, with a tiny loss in parameters.

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.

> **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.

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 $\epsilon >$ poly$(1/k)$, first shown by [5]; 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.

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 $x \in S$.*

Due to space limitations, we defer the analysis of derandomized DP tests to the full version of the paper. We remark that, in both independent and derandomized cases, we also get approximate, local, list-decoding algorithms for the corresponding DP codes.

1.4.2 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. For simplicity, here we consider CSPs with perfect completeness (i.e., with $\sigma = 1$); however, we can easily handle CSPs with imperfect completeness ($\sigma < 1$) by appropriately relaxing the acceptance condition in our PCPs.

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 always accepts an honest proof C_E for a 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 $k' = \Theta(\sqrt{k})$ 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 all constraints on each of the $B_{E,i}$'s[7], and
> *(b)* they assign the same values to A.

Theorem 1.4. *(i) If a CSP-instance (G, Φ) is satisfiable, then there is a proof C_E accepted by verifier $\mathcal{Y}$ with probability 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 $\epsilon = e^{-(1/c)\delta k'}$, for any δ, k' such that $k' > c/\delta$.*

The proof of this theorem is given in Section 4. Together with the PCP Theorem [2, 1] (e.g., using [3]), but *without* the parallel repetition theorem of [20], 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 4.1 in Section 4.

Before Raz's celebrated result [20], Feige and Kilian [8] and Verbitsky [23] 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 [8]. 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 [8].

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.

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

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 all 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 4.3 that $\mathcal{Y}'$ has soundness $\exp(-\delta k')$; in contrast, the analysis of [8] yields only inverse polynomial soundness.

Theorem 1.5. *(i) If a CSP-instance (G, Φ) is satisfiable, then there are proofs (C_1, C_E) accepted by verifier $\mathcal{Y}'$ with probability* 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 $\epsilon = e^{-(1/c)\delta k'}$, for any δ, k' such that $k' > c/\delta$.*

Currently, our rate of soundness decay $\exp(-\delta\sqrt{k})$ in Theorems 1.4 and 1.5 is never better than Rao's bound $\exp(-\delta^2 k)$ for projection games [19]. However, we see no reason why the soundness decay in our PCP constructions cannot be improved to $\exp(-\delta k)$.

2. PRELIMINARIES

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.

For our analysis of DP tests, we use basic sampling lemmas, which (as in [15]) 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 2.1 (Subset/Subspace Samplers). *Both Independent and Subspaces inclusion graphs $G(L, R)$ defined above are (α, β)-samplers, where*

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

The proof idea is to use the Chernoff-Hoeffding bound for Independent, and the Chebyshev bound for Subspaces. We omit the details due to space limitations.

We will also need the following basic properties of samplers; the proofs are straightforward.

Lemma 2.2. *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 $\left|\mathbf{Pr}_{r \in R', \ell \in N(r)}[\ell \in L'] - \lambda\right| \leqslant (2/3)\lambda$.*

Lemma 2.3. *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 $\left|\mathbf{Pr}_{r \in N(\ell)}[r \in R'] - \rho\right| \leqslant (2/3)\rho$.*

For sampler graphs in the Independent case, we can show a tighter version of Lemma 2.3, as follows; the proof is by a careful application of the standard Chernoff bounds.

Lemma 2.4. *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 $O((\log 1/\rho)/k)$*[8] *fraction of vertices $\ell \in L$, we have $\left|\mathbf{Pr}_{r \in N(\ell)}[r \in R'] - \rho\right| \leqslant \nu\rho$.*

Corollary 2.5. *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 2.4. □

3. ANALYSIS OF THE DP 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.

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

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

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

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

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 [8, 5], 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 [5] 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 $\epsilon > \mathrm{poly}(1/k)$ (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 [5].

3.1 Excellence

Using arguments similar to those in [15], we get

Lemma 3.1. *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* (A_0, B_0) *is good with probability at least* $\epsilon/2$.

Lemma 3.2. $\mathbf{Pr}_{A_0,B_0}[(A_0, B_0)$ *is good but not* (α, γ)-*excellent*$] < \gamma'/\gamma$*, where* $\gamma' = e^{-\Omega(\alpha k')}$.

As an immediate corollary of Lemmas 3.1 and 3.2, we get

Corollary 3.3. *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 3.4. *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 3.4. 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

$$\mathbf{Pr}_{B \subseteq \mathcal{U} \setminus A_0}[B \in Cons\ \&\ C(A_0, B)|_B \overset{>\beta}{\neq} g(B)] > \nu', \qquad (1)$$

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 3.4, and re-state it as Lemma 3.6 below. Here, we prove the result for the Boolean case; in the full version of the paper, we show how to reduce the general case to the Boolean case.

Definition 3.5. *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.[9]

[9] 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 $\mathcal{U} \setminus A_0$.

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

Lemma 3.6. *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 $\alpha > \Omega((\ln 1/\epsilon)/(k/k'))$.*

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

$$\mathbf{Pr}_{B\subseteq\mathcal{U}}[B \in Cons \ \&\ C'(B) \overset{>\beta}{\neq} g(B)] > \nu', \tag{2}$$

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, by Lemma 2.4 and Corollary 2.5, we get the following two claims.

Claim 3.7. *For all but at most $O((\ln 1/\epsilon)/k)$ fraction of inputs $x \in \mathcal{U}$, we have $|Cons_x|/|\mathcal{B}_x| \geqslant \epsilon/6$.*

Claim 3.8. *For all but at most $O((\ln 1/\epsilon)/(k/k'))$ fraction of k'-subsets $E \subset \mathcal{U} \setminus A_0$, we have $|Cons_E|/|\mathcal{B}_E| \geqslant \epsilon/6$.*

Claim 3.9. *Let x be any input such that $|Cons_x|/|\mathcal{B}_x| \geqslant \epsilon/6$. Then, for all but at most $O((\ln 1/\epsilon)/(k/k'))$ fraction of k'-sets E containing x, $\mathbf{Pr}_{B\in Cons_E}[C'(B)|_x = g(x)] \geqslant 1/10$.*

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 2.5, 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 3.10. *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 3.7, 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 3.9 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 3.10 the following corollary.

Claim 3.11. *For $\delta = O((\ln 1/\epsilon)/(k/k'))$ as in Claim 3.10, let $\delta' = 10\delta$, and let $\delta'' = 0.1$ (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 $x \in E$, $\mathbf{Pr}_{B\in Cons_E}[C'(B)|_x = g(x)] \geqslant 0.1]$.*

Proof of Lemma 3.6. Let $\delta' = 10\delta$ and $\delta'' = 1/10$, for the δ in Claim 3.11. We get by Claims 3.8 and 3.11 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 $x \in E$, $\mathbf{Pr}_{B'\in Cons_E}[C'(B')|_x = g(x)] \geqslant 1/10$.

Now consider the following distribution of subsets E: pick a random k-subset B satisfying the event of Eq. (2), and then pick a random k'-subset E of B. By Lemmas 2.1 and 2.2, 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. (2), we get by Chernoff-Hoeffding that almost all[10] 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. (2), 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

$$\mathbf{Pr}_{B':E\subset B'}[B' \in Cons \ \&\ C'(B')|_E \overset{>\beta/40}{\neq} C'(B)|_E] \geqslant \epsilon/120. \tag{3}$$

Since, for a random B conditioned on satisfying the event of Eq. (2), at least 1/4 fraction of sets E satisfy Eq. (3), 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. □

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

3.3 Local agreement implies global agreement

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

Lemma 3.12. *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 $k \geqslant \Omega(k'^2)$.*

Proof sketch. 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 3.3, it is an excellent set. By Lemma 3.4, 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 $A_1 = S \backslash B_1$; 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 3.4, 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. □

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

Here we give a simpler proof of the following result of [5]. The same argument also yields Theorem 1.3.

Theorem 3.13 ([5]). *There is a constant $0 < \eta < 1$ such that, 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 $k^{-\eta}$ fraction of $x \in S$.*

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[11]. 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 [5] for sublinear success precisely precludes such disjoint excellent pairs.

Claim 3.14. *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 sketch. The proof follows by analyzing the following equivalent 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 $B'' = B \cup A_2$). □

Claim 3.15. *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 3.14. Let g_{A_i} be the plurality function over sets in $Cons_{A_i,B_i}$, for $i = 1, 2$. For $\gamma \ll \epsilon^7$, $\mathbf{Pr}[(A_i, B_i)$ 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 B'' being a random set in $Cons_{A_1,B_1}$, we get by Lemma 3.4 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 3.14, this implies that the following event happens with probability at least $\Omega(\epsilon^5)$: (A_i, B_i) is (α, γ)-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 conditions imply $g_{A_1}(B) \overset{\leqslant\beta'}{\neq} C(A_1, B'')|_B$ 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 $g_{A_1}(B) \overset{\leqslant 2\beta'}{\neq} g_{A_2}(B)$. □

Claim 3.16. *For $\Omega(\epsilon^5)$ 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 3.15 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 $\mathbf{Pr}_{x \in \mathcal{U}}[g_{A_1}(x) \neq g_{A_2}(x)] > 2\alpha'$. 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 3.13. By Claim 3.15 and averaging, there are $\Omega(\epsilon^5)$ pairs (A_1, B_1) where $\mathbf{Pr}_{A_2,B_2,B}[(A_2, B_2)$ is (α, γ)-excellent & $g_{A_1}(\mathcal{U}) \overset{\leqslant\alpha'}{\neq} g_{A_2}(\mathcal{U})] \geqslant \Omega(\epsilon^5)$, with A_2, B_2, B chosen as in the random experiment of Claim 3.15, 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 $\mathrm{poly}(\epsilon)$ 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

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

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 3.4 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. □

4. 2-QUERY PCPS

4.1 Proof of Theorem 1.4

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).

For part *(i)*, an honest proof C_E (based on some satisfying assignment for (G, Φ)) will be accepted with probability 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[12].

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 3.2, 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 $k' > c/\delta$.

3. (A, a) is excellent. By Lemma 3.4, there is $g = g_{A,a} : U \to \Sigma$,[13] 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:

$$\mathbf{Pr}[C_E(A_{E,2}, B_{E,2})|_A = a \& C_E(A_{E,2}, B_{E,2})|_B \overset{>40\alpha}{\neq} g(B)] < \epsilon^2, \tag{4}$$

with the probability being over $A_{E,2}, B_{E,2}, B$, 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. 4 above.

Finally, by the soundness assumption for (G, Φ), every assignment violates at least δ fraction of edge constraints in G. In particular, this holds for our 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.

4.2 A new parallel repetition theorem

Let $G(V, E)$ be a d-regular graph, and let $C : E \to 2^{\Sigma^2}$ be a set of edge constraints. Consider the game $T = T(G, C)$, where the verifier picks a *pair* of edges at random (from some distribution P), sends one edge to each prover, 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, they agree on its label.

[12] 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.

[13] 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.

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. But one can also the following 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; this is very similar to miss-match proof systems of [8]. In particular, the famous counterexamples of Feige and Verbitsky[9] and of Raz [21] don't seem to hold for such games. Indeed, we get the following.

Theorem 4.1. *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 $(1 - (v/m))^{\Omega(k)}$.*

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

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 $C_1(A_0) = C_2(A_0, B_1)|_{A_0}$.

Here we define $Cons_{A_0}$ 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 4.1. 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 4.1.

We conclude by remarking that while our current techniques stop at the exponent $\sqrt{k}$, we see no obvious obstacle to improving it to k, and proving possibility/impossibility of this is an 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 to [18].

Acknowledgments. We thank Irit Dinur, Ariel Gabizon, Oded Goldreich, Dana Moshkovitz, Ryan O'Donnell, and Anup Rao for their comments, and 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 [8], which led us to the proof of Theorem 1.5.

5. REFERENCES

[1] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *JACM*, 45(3):501–555, 1998.

[2] S. Arora and S. Safra. Probabilistic checking of proofs: A new characterization of NP. *JACM*, 45(1):70–122, 1998.

[3] I. Dinur. The PCP theorem by gap amplification. *JACM*, 54(3), 2007.

[4] I. Dinur. PCPs with small soundness error. *ACM SIGACT News*, 2008.

[5] I. Dinur and E. Goldenberg. Locally testing direct products in the low error range. In *FOCS'08*.

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

[7] F. Ergün, S. Kannan, R. Kumar, R. Rubinfeld, and M. Viswanathan. Spot-checkers. *JCSS*, 60(3):717–751, 2000.

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

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

[10] O. Goldreich and L.A. Levin. A hard-core predicate for all one-way functions. In *STOC'89*, pages 25–32, 1989.

[11] O. Goldreich, N. Nisan, and A. Wigderson. On Yao's XOR-Lemma. *ECCC*, TR95-050, 1995.

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

[13] T. Holenstein. Parallel repetition: Simplifications and the no-signaling case. In *STOC'07*, pages 411–419, 2007.

[14] R. Impagliazzo, R. Jaiswal, and V. Kabanets. Approximately list-decoding direct product codes and uniform hardness amplification. In *FOCS'06*, pages 187–196, 2006.

[15] R. Impagliazzo, R. Jaiswal, V. Kabanets, and A. Wigderson. Uniform direct-product theorems: Simplified, optimized, and derandomized. In *STOC'08*, pages 579–588, 2008.

[16] R. Impagliazzo and A. Wigderson. P=BPP if E requires exponential circuits: Derandomizing the XOR Lemma. In *STOC'97*, pages 220–229, 1997.

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

[18] D. Moshkovitz and R. Raz. Two query PCP with sub-constant error. In *FOCS'08*, 2008.

[19] A. Rao. Parallel repetition in projection games and a concentration bound. In *STOC'08*, pages 1–10, 2008.

[20] R. Raz. A parallel repetition theorem. *SICOMP*, 27(3):763–803, 1998.

[21] R. Raz. A counterexample to strong parallel repetition. In *FOCS'08*, 2008.

[22] L. Trevisan. List-decoding using the XOR lemma. In *FOCS'03*, pages 126–135, 2003.

[23] O. Verbitsky. Towards the parallel repetition conjecture. *TCS*, 157(2):277–282, 1996.

[24] A.C. Yao. Theory and applications of trapdoor functions. In *FOCS'82*, pages 80–91, 1982.