Tolerant Linearity Testing and Locally Testable Codes

Swastik Kopparty[1] and Shubhangi Saraf[2]

[1] CSAIL, MIT, Cambridge MA 02139, USA
swastik@mit.edu
[2] CSAIL, MIT, Cambridge MA 02139, USA
shibs@mit.edu

Abstract. We study tolerant linearity testing under general distributions. Given groups G and H, a distribution μ on G, and oracle access to a function $f : G \to H$, we consider the task of approximating the smallest μ-distance of f to a homomorphism $h : G \to H$, where the μ-distance between f and h is the probability that $f(x) \neq h(x)$ when x is drawn according to the distribution μ. This question is intimately connected to local testability of linear codes.

In this work, we give a general sufficient condition on the distribution μ for linearity to be tolerantly testable with a constant number of queries. Using this condition we show that linearity is tolerantly testable for several natural classes of distributions including low bias, symmetric and product distributions. This gives a new and simple proof of a result of Kaufman and Sudan which shows that sparse, unbiased linear codes over $\mathbb{Z}_2^n$ are locally testable.

1 Introduction

Let $\mathcal{C}$ be a class of functions from a finite set $\mathcal{D}$ to a finite set $\mathcal{R}$. In the task of *tolerant testing* for $\mathcal{C}$, we are given oracle access to a function $f : \mathcal{D} \to \mathcal{R}$, and we wish to determine using few queries to f, whether f is well approximable by functions in $\mathcal{C}$; equivalently, to distinguish between the case when f is *close* to some element of $\mathcal{C}$, and the case when f is *far* from all elements of $\mathcal{C}$. Tolerant property testing was introduced by Parnas, Ron and Rubinfeld in [PRR06] as a refinement of the problem of property testing [RS96], [GGR98] (where one wants to distinguish the case of f *in* $\mathcal{C}$ from the case when f is *far* from $\mathcal{C}$), and is now widely studied. The usual notion of closeness considered in the literature is via the distance measure $\Delta(f, g) = \Pr_{x \in \mathcal{D}}[f(x) \neq g(x)]$, where $x \in \mathcal{D}$ is picked according to the *uniform distribution* over $\mathcal{D}$.

We propose to study tolerant property testing under general distributions. Given a probability measure μ on $\mathcal{D}$, the μ-distance of f from g, where $f, g : \mathcal{D} \to \mathcal{R}$, is defined by

$$\Delta_\mu(f, g) = \Pr_{x \in \mu}[f(x) \neq g(x)].$$

Then the measure of how well f can be approximated by elements of $\mathcal{C}$ is via the μ-distance

$$\Delta_\mu(f, \mathcal{C}) = \min_{g \in \mathcal{C}} \Delta_\mu(f, g).$$

The new goal in this context then becomes to approximate $\Delta_\mu(f, \mathcal{C})$ using only a few oracle calls to f. In this paper, we study a concrete instance of the above framework. We consider the original problem considered in the area of property testing, namely the classical problem of *linearity testing*.

The problem of linearity testing was introduced by Blum, Luby and Rubinfeld in [BLR93]. In this problem, we are given oracle access to a function $f : G \to H$, where G and H are abelian groups, and want to distinguish between the case that f is a *homomorphism* from G to H and the case that f is *far* from the class $\mathcal{C} = \mathsf{Hom}(G, H)$, of all homomorphisms from G to H. [BLR93] gave a simple 3-query test T that achieves this. In fact, this test also achieves the task of *tolerant linearity testing*; i.e., for any function $f : G \to H$, letting $\delta = \Pr[T^f \text{ rejects}]$, we have

$$C_1 \cdot \delta \leq \Delta_{U_G}(f, \mathsf{Hom}(G, H)) \leq C_2 \cdot \delta,$$

where C_1 and C_2 are absolute constants, and U_G is the uniform distribution on G. Hence the test of [BLR93], in addition to *testing* linearity, actually estimates how well f can be *approximated* by functions in $\mathcal{C} = \mathsf{Hom}(G, H)$.

Here we initiate the study of tolerant linearity testing over general probability distributions. Let μ be a probability distribution on an abelian group G. In the problem of *tolerant linearity testing under* μ, we wish to estimate the how well f may be approximated under μ by homomorphisms from G to H. For a given family $(G_n, H_n, \mu_n)_n$, we say *linearity is tolerantly testable* under $\mu = \mu_n$ with q queries, if there exists a q-query tester T_n and constants C_1, C_2 such that for any $f : G_n \to H_n$, letting $\delta = \Pr[T_n^f \text{ rejects}]$, we have

1. **Perfect completeness:** $\delta = 0$ if and only if $\Delta_\mu(f, \mathsf{Hom}(G_n, H_n)) = 0$.
2. **Distance approximation:** δ approximates $\Delta_\mu(f, \mathsf{Hom}(G_n, H_n))$:

$$C_1 \cdot \delta \leq \Delta_\mu(f, \mathsf{Hom}(G_n, H_n)) + o_n(1) \leq C_2 \cdot \delta. \tag{1}$$

We argue that this is indeed a natural definition under which to study linearity testing under general distributions. For one, this definition ensures that any "useful" queries made by the tester essentially have to be distributed according to μ. Without the "tolerant" aspect of the definition, a tester could potentially get access to "advice" by querying f at locations where μ has no support. For example, consider a scenario where f is given by a black box that runs in expected polynomial time under the distribution μ. In this setting, it is meaningful to ask how well f is approximated by linear functions under μ, although it is not as reasonable to expect access to evaluations of f at points not distributed according to μ. More importantly, the tolerant aspect of this definition give it a strong connection to locally testable codes (which we discuss shortly).

The problem of linearity testing (without the tolerant aspect) was studied in the setting of general distributions by [HK07]. They gave a simple 3-query

test that, given oracle access to the function $f : G \to H$, distinguishes between $f \in \mathsf{Hom}(G, H)$ and f that are far from $\mathsf{Hom}(G, H)$. In fact this tester does not even require an explicit description of μ, it simply requires access to samples from μ![3] However, unlike the BLR test, the [HK07] test is intolerant: the queries it makes are not according to the distribution μ.

The main question is to determine for which μ is linearity tolerantly testable under μ. This seems to be a basic question worthy of further study. Furthermore, the notion of linearity being tolerantly testable under general distributions is intimately connected with the concept of locally testable linear codes [GS02], and we now elaborate on this connection.

Connection to locally testable codes: Let $C \subseteq \mathbb{Z}_2^N$ be a linear code (we restrict to binary codes in this discussion). C is called *locally testable* if there is a constant query tester, that given oracle access to any $r : [N] \to \mathbb{Z}_2$, distinguishes between the case that $r \in C$ and r being far from C (in Hamming distance).

Now let C be any linear code, and let $s_1, \ldots, s_N \in \mathbb{Z}_2^n$ be the columns of a generator matrix for C. Let μ be the uniform distribution over $\{s_1, \ldots, s_N\}$. Then, if linearity is tolerantly testable under μ, then C is locally testable. Indeed, given any $r : [N] \to \mathbb{Z}_2$, we may define the function $f : \mathbb{Z}_2^n \to \mathbb{Z}_2$ by $f(x) = r(j)$ if $x = s_j$, and $f(x) = 0$ otherwise. By the tolerant testability of linearity under μ, any useful query made by a tolerant linearity tester for μ must be to one of the s_j. The distance of f from linear under μ then translates directly into the Hamming distance of r from C, and the very same tester that tolerantly tests linearity under μ shows that C is locally testable.

A more general goal behind the study of linearity testing under general distributions is to develop a better understanding what makes a code locally testable. We also believe that the theory of property testing under nonuniform distributions for other classes of functions $\mathcal{C}$ will be a fruitful and enlightening pursuit.

1.1 Main Notions and Results

Our main contribution is to highlight a simple criterion, which we call *uniform-correlatability*, that lets us design and analyze tolerant linearity tests under a given distribution. Roughly speaking, a distribution μ over an abelian group G is uniformly-correlatable if one can "design" a distribution of small random matrices with entries from G with each entry of the matrix distributed according to μ, while all the row-sums and column-sums are nearly uniformly distributed. In this case, we show that linearity is tolerantly testable under μ with few queries (see Theorem 1). We complement this by demonstrating that many natural distributions satisfy this criterion (see Theorems 2, 3, 4).

Definition 1 (Uniformly-correlatable distribution). *Let μ be a probability distribution on an abelian group G. We say that μ is* (ϵ, k)-uniformly-correlatable *if there is a random variable $\mathbf{X} = (\mathrm{X}_{ij})_{i,j \in [k]}$ taking values in $G^{k \times k}$ such that:*

[3] Tolerant linearity testing, however, necessarily requires more information about μ.

1. *For each $i, j \in [k]$, $\mathrm{X}_{i,j}$ is distributed according to μ.*
2. *For $i \in [k]$, let Y_i be the random variable $\sum_{j \in [k]} \mathrm{X}_{ij}$. For $j \in [k]$, Z_j be the random variable $\sum_{i \in [k]} \mathrm{X}_{ij}$. Then the distribution of $((\mathrm{Y}_i)_{i \in [k]}, (\mathrm{Z}_j)_{j \in k})$ is ϵ-close to the uniform distribution over the set $\{((y_i)_{i \in [k]}, (z_j)_{j \in [k]}) \in G^{2k} \mid \sum_{i \in [k]} y_i = \sum_{j \in [k]} z_j\}$.*

Our main result, given below, is that uniformly correlatable distributions are tolerantly testable.

Theorem 1 (Uniformly correlatable distributions are tolerantly testable). *Let μ be a distribution over G that is (ϵ, k)-uniformly-correlatable. Then there is a $4k$ query tester T such that for any $f : G \to H$, letting $\delta = \Pr[T^f \text{ rejects}]$, we have:*

1. **Perfect completeness:** *$\delta = 0$ if and only if $\Delta_\mu(f, \mathsf{Hom}(G, H)) = 0$*
2. **Distance approximation:**

$$\frac{\delta - 4\epsilon}{4k} \leq \Delta_\mu(f, \mathsf{Hom}(G, H)) \leq \frac{6k}{1 - 12\epsilon k} \cdot \delta.$$

Thus, for any μ which is (ϵ, k)-uniformly-correlatable for constant k and $\epsilon = o(1)$, we conclude that linearity is tolerantly testable under μ (in the sense of Equation (1)). We supplement the above theorem with following results, showing that some general classes of μ are all (ϵ, k)-uniformly-correlatable for suitable ϵ, k, and thus showing that linearity is tolerantly testable under all such μ.

Theorem 2 (Low-bias distributions are uniformly correlatable). *Let μ be a probability distribution on G such that for all nontrivial characters $\chi : G \to \mathbb{C}^\times$,*

$$|\mathop{\mathbf{E}}_{x \in \mu}[\chi(x)]| < |G|^{-\gamma}.$$

Then for $k = \Omega(1/\gamma)$, μ is $(|G|^{-\Omega(k\gamma)}, k)$-uniformly-correlatable.

Using the above theorem, we conclude that linearity is tolerantly testable under any low-bias distribution.

Corollary 1 (Low-bias distributions are tolerantly testable). *Let G be an abelian group, and let μ and γ be as in Theorem 2. Then there are constants $C_1 = C_1(\gamma)$, $C_2 = C_2(\gamma)$, and a $O(1/\gamma)$-query test T such that for any abelian group H and any function $f : G \to H$, letting $\delta = \Pr[T^f \text{ rejects}]$, we have*

$$C_1\delta \leq \Delta_\mu(f, \mathsf{Hom}(G, H)) + o(1/|G|) \leq C_2\delta.$$

For the special case of $G = \mathbb{Z}_2^n$, $H = \mathbb{Z}_2$, and μ being the uniform distribution over some set, via the connection described in Section 1, this gives a new proof of a result of Kaufman and Sudan [KS07], who proved that sparse, low-bias linear codes are locally testable (in particular, that sparse random linear codes are locally testable). Their proof used the machinery of Krawtchouk polynomials

and nontrivial information about the distribution of their roots. The corollary above gives a new and simple proof of this fact[4], and generalizes it to arbitrary G and H.

In the next theorem, we show that product distributions over $\mathbb{Z}_2^n$ are uniformly correlatable, whenever the individual distributions are not too biased. We then use our main theorem to conclude that linearity is tolerantly testable under such distributions. The proof is by reducing to the $n = 1$ case, and is omitted in this version of the paper.

Theorem 3 (Product Distributions are uniformly correlatable). *Let $G = \mathbb{Z}_2^n$ and let $p_1, \ldots, p_n \in [\gamma, (1-\gamma)]$. For each $i \in [n]$, let μ_i be the distribution over $\mathbb{Z}_2$ with $\mu_i(1) = p_i$, Let μ be the product distribution $\prod_{i=1}^n \mu_i$ on G. Then μ is $(0, O(1/\gamma))$-uniformly-correlatable.*

Corollary 2 (Product Distributions are tolerantly testable). *Let μ and γ be as in Theorem 3. Then there are constants $C_1 = C_1(\gamma)$, $C_2 = C_2(\gamma)$, and a $O(1/\gamma)$-query test T such that for any abelian group H and any function $f : G \to H$, letting $\delta = \Pr[T^f \text{ rejects}]$, we have*

$$C_1\delta \leq \Delta_\mu(f, \mathsf{Hom}(G, H)) \leq C_2\delta.$$

In the next theorem, we consider distributions on $\mathbb{Z}_2^n$ that are symmetric under permutations of the coordinates. For technical reasons, we only show correlatability for distributions supported on words of even Hamming weight, but this suffices to show testability for general symmetric distributions μ. The proof is by a volume growing argument, and is omitted in this version of the paper.

Theorem 4 (Symmetric distributions are uniformly correlatable). *Let $G' = \mathbb{Z}_2^n$ and let G be the subgroup of G' consisting of even weight words. Let μ be a distribution on G, symmetric under permutations of the coordinates of G', and supported on words whose weights lie in the interval $[\gamma n, (1-\gamma)n]$. Then μ (viewed as a distribution on G) is $(0, O(1/\gamma))$-uniformly-correlatable.*

Corollary 3 (Symmetric distributions are tolerantly testable). *Let μ be a symmetric distribution on $G = \mathbb{Z}_2^n$ such that supported on words whose weights lie in the interval $[\gamma n, (1-\gamma)n]$. Then there are constants $C_1 = C_1(\gamma)$, $C_2 = C_2(\gamma)$, and a $O(1/\gamma)$-query test T such that for any abelian group H and any function $f : G \to H$, letting $\delta = \Pr[T^f \text{ rejects}]$, we have*

$$C_1\delta \leq \Delta_\mu(f, \mathsf{Hom}(G, H)) \leq C_2\delta.$$

1.2 Other Related Work

Kiwi [Kiw03] considered puncturings of the Hadamard code and gave a sufficient condition for certain codes to be locally testable. There has been a large body

[4] The $o(1/|G|)$ term in this corollary, combined with the perfect completeness of the test, in fact shows that the corresponding code is "strongly" locally testable.

of work constructing short locally testable codes [GS02], [GR05], [BSSVW03], [BSGH+04], [BSS05], [Din06], [Mei08]. In our framework, these correspond to distributions μ over $\mathbb{Z}_2^n$ supported on sets of size poly(n) under which linearity is tolerantly testable. It would be interesting to obtain a better understanding of which μ with such small support/min-entropy are such that linearity is tolerantly testable under μ.

Property testing under nonuniform distributions has arisen naturally and studied in several other contexts (in addition to [HK07]). The problem of dictatorship testing under the p-biased distribution arose in the work of Dinur and Safra [DS02] on the inapproximability of VERTEX-COVER. Subsequently, the problem of junta-testing [KS03] was also considered under the p-biased distribution.

In [AKK+03], Alon et. al. gave a constant query test for low degree polynomials over $\mathbb{F}_2$ (this was later extended to larger fields by [KR04], [JPRZ04], [KS08]). A suitable application of this test shows that for any p of the form $\frac{a}{2^b}$ for constant b, there is a constant query test for low degree polynomials under the p-biased distribution. This lends optimism to the goal of understanding the more general question of testing low degree polynomials under general distributions.

Paper organization: In Section 2, we give an overview of our proofs. In Section 3, we prove that uniformly-correlatable distributions are tolerantly testable (Theorem 1). The proof of Theorem 2 appears in Section 4. Finally in Section 5, we discuss some problems and directions for further study.

2 Overview of Proofs

We first give some intuition for the uniform correlatability criterion. For T to be a tester for linearity under μ, it needs to satisfy the following minimum requirements: (1) each query made by the tester needs to be distributed essentially according to μ (so that the probability of rejection is upper bounded by the distance), and (2) the queries need to satisfy some linear relations (so that the tester has something to test). This already indicates that a tester will need to "design" a query distribution very carefully, so that both the above requirements are satisfied. This is where the uniformly-correlatable criterion comes in: given the uniformly-correlated distribution on matrices, it allows us to design other correlations quite flexibly, and in particular to produce queries distributed according to μ that satisfy linear relations.

The proof of Theorem 1 follows the rough outline of the original "self-correction" proof of the BLR linearity test (for linearity testing under the uniform distribution)[5]. It proceeds in 3 steps: we first define a "self-corrected" version of the function being tested and show that the function being tested is μ-close to that function. We then show that the self-corrected version is in fact self-corrected with overwhelming probability. Finally we use the above two

[5] Note that the uniform distribution is $(0, 1)$-uniformly-correlatable, and for this case, the test given by Theorem 1 essentially reduces to the BLR linearity test.

facts to show that the self-corrected function is in fact a homomorphism. We use the correlated matrix (given by uniform correlatability) to construct two tests: each of these tests helps with a particular step of the proof. In contrast, the BLR [BLR93] linearity test makes only one kind of test which miraculously suffices for all steps of the analysis.

We show Theorem 2 on uniform-correlatability of low-bias distributions by Fourier analysis, using a version of the Vazirani-XOR lemma. For Theorem 3, we use the closure property of uniform-correlatability under products, and it thus suffices to show that the p-biased distribution on $\mathbb{Z}_2$ is uniformly correlatable. We then exhibit such a correlation by a direct construction. Finally, to show Theorem 4, we use the closure property of uniform-correlatability under convex combinations. This reduces the question to showing that for any even $w \in [\gamma n, (1-\gamma)n]$, the uniform distribution on vectors in $\mathbb{Z}_2^n$ of Hamming weight exactly w is uniformly correlatable (to get a uniform distribution on all words of even weight). This is a technically involved step, and is achieved by carefully analyzing the set of possible row-sums and column-sums of matrices with all entries being words of weight w. The correlatability of symmetric distributions supported on even weight words, along with an additional trick, then allows us to deduce that linearity is tolerantly testable under any symmetric distribution over words with weights in $[\gamma n, (1-\gamma)n]$.

3 Uniformly Correlatable Distributions are Testable

Let μ be (ϵ, k)-uniformly-correlatable. Fix a distribution μ^{mat} over $G^{k\times k}$ witnessing this property. Without loss of generality, we may assume that all the rows and columns of μ^{mat} are identically distributed, and let μ^{row} be this distribution (indeed, we may take a random sample from μ^{mat}, randomly permute the rows and columns, and then transpose it with probability $\frac{1}{2}$: the distribution of the resulting matrix witnesses the correlatability property, and also has identical row and column distributions). We now define a few distributions related to it:

1. **Distribution $\mu^{\mathrm{mat}}_{(r,s)}$:** For $r, s \in G^k$ with $\sum_{i\in[k]} r_i = \sum_{j\in[k]} s_j$, $\mu^{\mathrm{mat}}_{(r,s)}$ is the distribution of samples X from μ^{mat} conditioned on $\sum_{j\in[k]} X_{ij} = r_i$ and $\sum_{i\in[k]} X_{ij} = s_j$.
2. **Distribution μ^{row}_r:** For $r \in G$, μ^{row}_r is the distribution of samples x from μ^{row} conditioned on $\sum_{i\in[k]} x_i = r$.
3. **Distribution μ^*_r:** This is the distribution of the random variables $(y, z) \in G^k \times G^k$ produced by the following random process. Let U_G be the uniform distribution on G. First sample r' from U_G. Then independently sample y from $\mu^{\mathrm{row}}_{r+r'}$ and z from $\mu^{\mathrm{row}}_{r'}$. In particular, $\sum_i y_i - \sum_j z_j = r$.

We may now describe the linearity test under μ.

Test T: With probability 1/2, perform Test T1, and with probability 1/2 perform Test T2.

- **Test T1:** Sample r from μ. Sample $(y, z) \in G^{2k}$ from μ^*_r. If $\sum_{i\in[k]} f(y_i) - \sum_{i\in[k]} f(z_i) = f(r)$, then accept, else reject.

– **Test T2:** Sample r from U_G. Independently sample (y, y') and (z, z') from μ^*_r. If $\sum_{i\in[k]} f(y_i) - \sum_{i\in[k]} f(y'_i) = \sum_{i\in[k]} f(z_i) - \sum_{i\in[k]} f(z'_i)$, then accept, else reject.

It is clear that this test has perfect completeness.

The following fact basic fact about distributions will be useful while analyzing the test.

Fact 5 *Let R, S be random variables, and let h be function such that the distribution of $h(R)$ is ϵ-close to the distribution of S. Consider the distribution of R' sampled as follows: first pick S, and then let R' be a sample of R conditioned on $h(R) = S$. Then the distribution of R' is ϵ-close to the distribution of R.*

We now prove that the Test T is indeed a tester for linearity under μ, hence completing the proof of Theorem 1.

Theorem 6. *Let $f : G \to H$ and let $\delta \stackrel{\text{def}}{=} \Pr[T^f \text{ rejects}]$. Then,*

$$\frac{\delta - 4\epsilon}{4k} \leq \Delta_\mu(f, \mathsf{Hom}(G, H)) \leq \frac{12k}{1 - 24\epsilon k} \cdot \delta.$$

Proof. Let $f : G \to H$ and let $\delta = \Pr[T^f \text{ rejects}]$. Notice that by Fact 5, each k-tuple of queries made by the test T is ϵ-close to the distribution μ^{row}. Therefore, the probability that no query is made to an element of G where f disagrees with its nearest homomorphism in $\mathsf{Hom}(G, H)$ is at most $4k \cdot \Delta_\mu(f, \mathsf{Hom}(G, H)) + 4\epsilon$. Thus $\delta \leq 4k \cdot \Delta_\mu(f, \mathsf{Hom}(G, H)) + 4\epsilon$, which is the first inequality.

We now show the second inequality. If $\delta \geq \frac{1}{12k} - 2\epsilon$, then the claim is trivial (since $\Delta_\mu(\cdot, \cdot) \leq 1$). Suppose $\delta \leq \frac{1}{12k} - 2\epsilon$. Let δ_1 be the probability that Test T1 rejects. Let δ_2 be the probability that Test T2 rejects. Then $\delta = \frac{1}{2}(\delta_1 + \delta_2)$.

For $x \in G$, define the "self-corrected" value $g(x)$ to be the most probable value of $\sum_{i\in[k]} f(y_i) - \sum_{i\in[k]} f(z_i)$, where $(y, z) \in \mu^*_x$.

Lemma 1 (g is close to f). $\Delta_\mu(f, g) < 2\delta_1$.

Proof. Let $B = \{x \in G : g(x) \neq f(x)\}$.

For any $x \in G$, define

$$p_x = \Pr_{(y,z)\in\mu^*_x}\left[\sum_{i\in[k]} f(y_i) - \sum_{j\in[k]} f(z_j) \neq f(x)\right].$$

Notice that for any $x \in B$, $p_x \geq 1/2$. By definition, $\delta_1 = \mathbf{E}_{x\in\mu}[p_x]$. Applying Markov's inequality, we conclude that

$$\Pr_{x\in\mu}[x \in B] \leq \Pr_{x\in\mu}[p_x \geq 1/2] \leq 2\delta_1.$$

We now show that g is in fact a homomorphism.

Lemma 2 (Majority votes of g are overwhelming majorities). *For all $x \in G$,*

$$\Pr_{(y,z)\in\mu_x^*}[\sum_{i\in[k]} f(y_i) - \sum_{j\in[k]} f(z_j) \neq g(x)] \leq 2\delta_2.$$

Proof. Let $x \in G$. Take two independent samples (y^1, z^1) and (y^2, z^2) from μ_x^*. We will show that

$$\Pr\left[\sum_{i\in[k]} f(y_i^1) - \sum_{j\in[k]} f(z_j^1) \neq \sum_{i\in[k]} f(y_i^2) - \sum_{j\in[k]} f(z_j^2)\right] \leq 2\delta_2. \tag{2}$$

The lemma follows immediately from this.

We now prove Equation 2. By definition, (y^1, z^1) was generated by picking $r^1 \in U_G$, and then picking $y^1 \in \mu^{\text{row}}_{r^1+x}$ and $z^1 \in \mu^{\text{row}}_{r^1}$. Similarly (y^2, z^2) was generated by picking $r^2 \in U_G$, and then picking $y^2 \in \mu^{\text{row}}_{r^2+x}$ and $z^2 \in \mu^{\text{row}}_{r^2}$. Observe that both (y^1, y^2) and (z^1, z^2) come from the distribution $\mu^*_{r^1-r^2}$ (albeit not independently). Let (w^1, w^2) be another sample from $\mu^*_{r^1-r^2}$ (independent of (y^1, y^2) and (z^1, z^2)).

We now rewrite and then bound the left hand side of Equation (2) by:

$$\begin{aligned}
&\Pr\left[\sum_{i\in[k]} f(y_i^1) - \sum_{j\in[k]} f(z_j^1) \neq \sum_{i\in[k]} f(y_i^2) - \sum_{j\in[k]} f(z_j^2)\right] \\
=&\Pr\left[\sum_{i\in[k]} f(y_i^1) - \sum_{i\in[k]} f(y_i^2) \neq \sum_{j\in[k]} f(z_j^1) - \sum_{j\in[k]} f(z_j^2)\right] \\
\leq&\Pr\left[\sum_{i\in[k]} f(y_i^1) - \sum_{i\in[k]} f(y_i^2) \neq \sum_{j\in[k]} f(w_j^1) - \sum_{j\in[k]} f(w_j^2)\right] \\
&\qquad + \Pr\left[\sum_{i\in[k]} f(z_i^1) - \sum_{i\in[k]} f(z_i^2) \neq \sum_{j\in[k]} f(w_j^1) - \sum_{j\in[k]} f(w_j^2)\right]
\end{aligned}$$

Finally, note that $r^1 - r^2$ is uniformly distributed over G. Since $(y_1, y_2), (w_1, w_2)$ are independent samples from $\mu^*_{r_1-r_2}$ (and similarly for $(z_1, z_2), (w_1, w_2)$), this implies that both the terms in the last expression above equal the rejection probability of Test T2 ($= \delta_2$). This completes the proof of the lemma.

Lemma 3 (g is linear). $g \in \mathsf{Hom}(G, H)$.

Proof. Pick any $x, x' \in G^k$, and let $t = \sum_{i=1}^k x_i - \sum_{i=1}^k x'_i$. We will show that $\sum_{i=1}^k g(x_i) - \sum_{i=1}^k g(x'_i) = g(t)$.

We now describe a random process. Pick $(\alpha, \beta) \in \mu_t^*$. Pick $r \in U_G$. Pick r^1, s^1, r^2, s^2 uniformly from G^k conditioned on $\sum_{i\in[k]} r_i^1 = \sum_{i\in[k]} s_i^1 = r$, $\sum_{i\in[k]} r_i^2 = r - \sum_{i\in[k]} \alpha_i$, and $\sum_{i\in[k]} s_i^2 = r - \sum_{i\in[k]} x_i$.

Now pick random matrices $A \in \mu^{\mathrm{mat}}_{(r^1,s^1)}$, $A' \in \mu^{\mathrm{mat}}_{(r^2,s^1-\alpha)}$, $B \in \mu^{\mathrm{mat}}_{(r^1-x,s^2)}$, and $B' \in \mu^{\mathrm{mat}}_{(r^2-x,s^2-\beta)}$.

$$\begin{array}{cccc|cccc|c}
A_{11} & A_{12} & \cdots & A_{1k} & A'_{11} & A'_{12} & \cdots & A'_{1k} & \alpha_1 \\
A_{21} & A_{22} & \cdots & A_{2k} & A'_{21} & A'_{22} & \cdots & A'_{2k} & \alpha_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
A_{k1} & A_{k2} & \cdots & A_{kk} & A'_{k1} & A'_{k2} & \cdots & A'_{kk} & \alpha_k \\
\hline
B_{11} & B_{12} & \cdots & B_{1k} & B'_{11} & B'_{12} & \cdots & B'_{1k} & \beta_1 \\
B_{21} & B_{22} & \cdots & B_{2k} & B'_{21} & B'_{22} & \cdots & B'_{2k} & \beta_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
B_{k1} & B_{k2} & \cdots & B_{kk} & B'_{k1} & B'_{k2} & \cdots & B'_{kk} & \beta_k \\
\hline
x_1 & x_2 & \cdots & x_k & x'_1 & x'_2 & \cdots & x'_k & t
\end{array}$$

Arrange these random variables in a matrix, as shown in the figure.

First notice that by Fact 5, the distribution of each α_i is ϵ-close to the distribution μ. Similarly, the distribution of each β_i is ϵ-close to the distribution μ.

Let us study the row distribution. Again by Fact 5, for each $i \in [k]$, the distribution of $(A_{i\bullet}, A'_{i\bullet})$ is 4ϵ-close to $\mu^*_{\alpha_i}$. Similarly, for each $i \in [k]$, the distribution of $(B_{i\bullet}, B'_{i\bullet})$ is 4ϵ-close to $\mu^*_{\beta_i}$.

Now consider the distribution of the columns. For each $j \in [k]$, the distribution of $(A_{\bullet j}, B_{\bullet j})$ is 4ϵ-close to $\mu^*_{x_j}$. Similarly, for each $j \in [k]$, the distribution of $(A'_{\bullet j}, B'_{\bullet j})$ is 4ϵ-close to $\mu^*_{x'_j}$.

Summarizing, each row distribution $(A_{i\bullet}, A'_{i\bullet}, \alpha_i)$ and $(B_{i\bullet}, B'_{i\bullet}, \beta_i)$ is 5ϵ-close to the distribution of the distribution of queries of Test T1. Thus by a union bound, with probability at least $1-2k\cdot(\delta_1+5\epsilon)$, both the following events occur:

- **Event 1:** For each $i \in [k]$, $\sum_{j\in[k]} f(A_{ij}) - \sum_{j\in[k]} f(A'_{ij}) = f(\alpha_i)$,
- **Event 2:** For each $i \in [k]$, $\sum_{j\in[k]} f(B_{ij}) - \sum_{j\in[k]} f(B'_{ij}) = f(\beta_i)$.

Each column distribution $(A_{\bullet j}, B_{\bullet j})$ is 4ϵ-close to $\mu^*_{x_j}$, and each column distribution $(A'_{\bullet j}, B'_{\bullet j})$ is 4ϵ-close to $\mu^*_{x'_j}$. Thus by Lemma 2 and a union bound, with probability at least $1 - 2k \cdot (2\delta_2 + 4\epsilon)$, both the following events occur.

- **Event 3:** For each $j \in [k]$, $\sum_{i\in[k]} f(A_{ij}) - \sum_{i\in[k]} f(B_{ij}) = g(x_j)$,
- **Event 4:** For each $j \in [k]$, $\sum_{i\in[k]} f(A'_{ij}) - \sum_{i\in[k]} f(B'_{ij}) = g(x'_j)$,

Finally, since (α, β) was picked from μ^*_t, Lemma 2 tells us that the following event occurs with probability at least $1 - 2\delta_2$:

- **Event 5:** $\sum_{j\in[k]} f(\alpha_j) - \sum_{j\in[k]} f(\beta_j) = g(t)$,

Thus Events 1,2,3, 4 and 5 all occur with probability at least $1 - (2k+1)\cdot(\delta_1 + 2\delta_2 + 5\epsilon) > 0$, since we assumed that $\delta = \frac{\delta_1+\delta_2}{2} < \frac{1}{12k} - 2\epsilon$. In this case, we see

that

$$
\begin{aligned}
g(t) &= \sum_{i=1}^{k} f(\alpha_i) - \sum_{i=1}^{k} f(\beta_i) && \text{Event 5} \\
&= \sum_{i=1}^{k} \left(\sum_{j=1}^{k} (f(A_{ij}) - f(A'_{ij})) \right) - \sum_{i=1}^{k} \left(\sum_{j=1}^{k} (f(B_{ij}) - f(B'_{ij})) \right) && \text{Events 1 and 2} \\
&= \sum_{j=1}^{k} \left(\sum_{i=1}^{k} (f(A_{ij}) - f(B_{ij}))) \right) - \sum_{j=1}^{k} \left(\sum_{i=1}^{k} (f(A'_{ij}) - f(B'_{ij})) \right) && \text{rearranging terms} \\
&= \sum_{j=1}^{k} (g(x_j) - g(x'_j)) && \text{Events 3 and 4.}
\end{aligned}
$$

Hence, $\Pr[g(t) = \sum_{j=1}^{k}(g(x_j) - g(x'_j))] > 0$. However, this statement is a deterministic statement, and hence we conclude that $g(t) = \sum_{j=1}^{k}(g(x_j) - g(x'_j))$ Since this holds for every choice of x, x', g must be a homomorphism.

Thus, $\Delta_\mu(f, \mathsf{Hom}(G,H)) \leq \Delta_\mu(f,g) \leq 2\delta_1 \leq 4\delta \leq \frac{6k}{1-12\epsilon k} \cdot \delta$.

4 Low bias distributions are uniformly correlatable

In this section we prove Theorem 2.

Theorem 2 *Let μ be a probability distribution on G such that for all non-trivial characters $\chi : G \to \mathbb{C}^\times$,*

$$|\mathop{\mathbf{E}}_{x \in \mu}[\chi(x)]| < |G|^{-\gamma}.$$

Then for $k = \Omega(1/\gamma)$, μ is $(|G|^{-\Omega(k\gamma)}, k)$-uniformly-correlatable.

Proof. We begin with a lemma, which gives a simple criterion for checking that a distribution is close to uniform on a subgroup of G^t. It is an intermediate claim in the usual proof of the Vazirani XOR lemma [Gol95] which bounds the distance to uniform in terms of the maximum bias of the distribution. The full Vazirani XOR lemma turns out to be too weak for our purposes.

Lemma 4. *Let $S = \{(y_1, \ldots, y_t, z_1, \ldots, z_t) \in G^{2t} \mid \sum_{i\in[t]} y_i = \sum_{i\in[t]} z_i\}$. Let $(Y_1, \ldots, Y_t, Z_1, \ldots, Z_t)$ be an S-valued random variable. Suppose*

$$\sum_{\alpha_1,\ldots,\alpha_t,\beta_1,\ldots,\beta_t \in \hat{G} \text{ not all equal}} \left| \mathbf{E}\left[\prod_{i\in[t]} \chi_{\alpha_i}(Y_i) \cdot \prod_{j\in[t]} \overline{\chi_{\beta_j}(Z_j)} \right] \right|^2 \leq \lambda.$$

Then the distribution of $(Y_1, \ldots, Y_t, Z_1, \ldots, Z_t)$ is $\sqrt{\lambda}$-close to the uniform distribution over S.

We omit the proof of this lemma.

We can now prove the theorem. Let $\eta = |G|^{-\gamma}$. The distribution $\mathbf{X} = (\mathrm{X}_{ij})_{i,j\in[k]}$ is given by picking each X_{ij} independently from μ. For $i \in [k]$, let Y_i be the random variable $\sum_{j\in[k]} \mathrm{X}_{ij}$. For $j \in [k]$, Z_j be the random variable $\sum_{i\in[k]} \mathrm{X}_{ij}$. We wish to show that $(Y_1, \ldots, Y_k, Z_1, \ldots, Z_k)$ is $|G|^{-\Omega(k\gamma)}$-close to uniformly distributed on S.

In order to apply Lemma 4, we compute

$$\begin{aligned}
\sum_{\substack{\alpha_1,\ldots,\alpha_k\\ \beta_1,\ldots,\beta_k\\ \text{n.a.e.}}} \left|\mathbf{E}\left[\prod_{i\in[k]} \chi_{\alpha_i}(Y_i)\cdot \prod_{j\in[k]} \overline{\chi_{\beta_j}(Z_j)}\right]\right|^2 &= \sum_{\substack{\alpha_1,\ldots,\alpha_k\\ \beta_1,\ldots,\beta_k\\ \text{n.a.e.}}} \left(\mathbf{E}\left[\prod_{i,j} \chi_{\alpha_i-\beta_j}(\mathrm{X}_{ij})\right]\right)^2 \\
&= \sum_{\alpha,\beta} \left(\prod_{i,j} \mathbf{E}[\chi_{\alpha_i-\beta_j}(\mathrm{X}_{ij})]\right)^2 \\
&\qquad \text{(since the } \mathrm{X}_{ij} \text{ are independent)} \\
&\le \sum_{\alpha,\beta} \left(\eta^{|\{(i,j)\in[k]^2:\alpha_i\neq\beta_j\}|}\right)^2
\end{aligned}$$

Consider the term corresponding to $\alpha_1, \ldots, \alpha_k$ and $\beta_1, \ldots, \beta_k$. We classify the terms into 3 kinds, and separately bound the total contribution of terms of each kind.

Case A: the most frequently occurring element in $\alpha_1, \ldots, \alpha_k$ occurs at most $2k/3$ times. Then $|\{(i,j) \in [k]^2 : \alpha_i \neq \beta_j\}| \geq k \cdot k/3 = k^2/3$. Thus the sum of all terms in case A is at most $|G|^{2k} \cdot \eta^{2k^2/3}$.

Case B: the most frequently occurring element in $\alpha_1, \ldots, \alpha_k$ occurs at least $2k/3$ times, and that same element occurs in $\beta_1, \ldots, \beta_k$ at most $2k/3$ times. Then $|\{(i,j) \in [k]^2 : \alpha_i \neq \beta_j\}| \geq (k/3) \cdot (2k/3) = 2k^2/9$. Thus the sum of all terms in case B is at most $|G|^{2k} \cdot \eta^{2k^2/9}$.

Case C: Now suppose we are not in either of the above two cases. Suppose the most frequently occurring element in $\alpha_1, \ldots, \alpha_k$ occurs $a > 2k/3$ times, and that same element appears in $\beta_1, \ldots, \beta_k$ occurs $b > 2k/3$ times. Note that by the not all equal assumption, at most one of a, b can be equal to k. Then $|\{(i,j) \in [k]^2 : \alpha_i \neq \beta_j\}| \geq a \cdot (k-b) + b \cdot (k-a)$. Thus the total contribution of terms from Case C is at most (here we subtracted off the terms with $a = b = k$):

$$\sum_{a=2k/3}^{k} \sum_{b=2k/3}^{k} \binom{k}{a}\binom{k}{b} |G|^{k-a}|G|^{k-b}|G|\eta^{a(k-b)}\eta^{b(k-a)} - \binom{k}{k}\binom{k}{k}|G|.$$

This can be bounded from above by

$$|G| \cdot \left(\sum_{a=2k/3}^{k} \sum_{b=2k/3}^{k} \binom{k}{a}\binom{k}{b} |G|^{k-a}|G|^{k-b}\eta^{(k/2)(k-b)}\eta^{(k/2)(k-a)} - 1\right),$$

which in turn may be upper bounded by

$$|G| \cdot \Big((1 + |G|\eta^{(k/2)})^k (1 + |G|\eta^{(k/2)})^k - 1 \Big) \leq |G| \cdot (8k|G|\eta^{k/2}),$$

where the last inequality uses the fact that $k = \Omega(1/\gamma)$, and hence $\eta^{k/2} \ll (|G|k)^{-1}$. Summarizing, the sum of all the terms is at most $|G|^2\eta^{\Omega(k)} + |G|^{2k}\eta^{\Omega(k^2)}$. For $k = \Omega(1/\gamma)$, this quantity is at most $|G|^{-\Omega(\gamma k)}$. Lemma 4 now implies the desired result.

5 Discussion, Problems and Directions

We believe that there are many fruitful and interesting questions waiting to be explored in tolerant property testing under nonuniform distributions in general and tolerant linearity testing under nonuniform distributions in particular.

As far as we know, every distribution of linear min-entropy with bias at most 0.9 (say) is uniformly $(o(1), O(1))$-uniformly correlatable, and hence tolerantly testable. In fact, we do not even know of a single μ of linear min-entropy under which linearity is not tolerantly testable.

Question 1. Let μ be a probability distribution on $\mathbb{Z}_2^n$ with min-entropy $\Omega(n)$. Find necessary and sufficient conditions on μ for linearity to be tolerantly testable under μ.

Via the connection between tolerant linearity testing and local testability of codes, we even venture the following conjecture.

Conjecture 1. Every linear code $C \subseteq \mathbb{Z}_2^N$ with $N^{O(1)}$ codewords is locally testable!

Acknowledgements

We are very grateful to Madhu Sudan and Tali Kaufman for encouragement and invaluable discussions.

References

[AKK+03] Noga Alon, Tali Kaufman, Michael Krivelevich, Simon Litsyn, and Dana Ron. Testing low-degree polynomials over GF(2). In *Proceedings of RANDOM 2003, LNCS, vol. 2764*, pages 188–199, New York, 2003. Springer.

[BLR93] Manuel Blum, Michael Luby, and Ronitt Rubinfeld. Self-testing/correcting with applications to numerical problems. *Journal of Computer and System Sciences*, 47(3):549–595, 1993.

[BSGH+04] Eli Ben-Sasson, Oded Goldreich, Prahladh Harsha, Madhu Sudan, and Salil Vadhan. Robust PCPs of proximity, shorter PCPs and applications to coding. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing*, pages 1–10, New York, 2004. ACM Press.

[BSS05] Eli Ben-Sasson and Madhu Sudan. Short PCPs with poly-log rate and query complexity. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 266–275, New York, 2005. ACM Press.

[BSSVW03] Eli Ben-Sasson, Madhu Sudan, Salil Vadhan, and Avi Wigderson. Randomness efficient low-degree tests and short PCPs via ϵ-biased sets. In *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pages 612–621, New York, 2003. ACM Press.

[Din06] Irit Dinur. The PCP theorem by gap amplification. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 241–250, New York, 2006. ACM Press. Preliminary version appeared as an ECCC Technical Report TR05-046.

[DS02] Irit Dinur and Shmuel Safra. The importance of being biased. In *Proceedings on 34th Annual ACM Symposium on Theory of Computing*, pages 33–42, Montreal, Quebec, Canada, 19-21 May 2002.

[GGR98] Oded Goldreich, Shari Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. *JACM*, 45(4):653–750, 1998.

[Gol95] Oded Goldreich. Three xor-lemmas - an exposition. *Electronic Colloquium on Computational Complexity (ECCC)*, 2(56), 1995.

[GR05] Venkatesan Guruswami and Atri Rudra. Tolerant locally testable codes. In *APPROX-RANDOM*, pages 306–317, 2005.

[GS02] Oded Goldreich and Madhu Sudan. Locally testable codes and PCPs of almost-linear length. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, Vancouver, Canada, 16-19 November 2002.

[HK07] Shirley Halevy and Eyal Kushilevitz. Distribution-free property-testing. *SIAM J. Comput.*, 37(4):1107–1138, 2007.

[JPRZ04] Charanjit S. Jutla, Anindya C. Patthak, Atri Rudra, and David Zuckerman. Testing low-degree polynomials over prime fields. In *FOCS '04: Proceedings of the Forty-Fifth Annual IEEE Symposium on Foundations of Computer Science*, pages 423–432. IEEE Computer Society, 2004.

[Kiw03] Marcos A. Kiwi. Algebraic testing and weight distributions of codes. *Theor. Comput. Sci.*, 1-3(299):81–106, 2003.

[KR04] T. Kaufman and D. Ron. Testing polynomials over general fields. In *Proceedings of the Forty-fifthth Annual Symposium on Foundations of Computer Science*, pages 413–422, 2004.

[KS03] Guy Kindler and Shmuel Safra. Noise-resistant boolean-functions are juntas, April 17 2003.

[KS07] Tali Kaufman and Madhu Sudan. Sparse random linear codes are locally decodable and testable. In *FOCS*, pages 590–600. IEEE Computer Society, 2007.

[KS08] Tali Kaufman and Madhu Sudan. Algebraic property testing: the role of invariance. In Richard E. Ladner and Cynthia Dwork, editors, *STOC*, pages 403–412. ACM, 2008.

[Mei08] Or Meir. Combinatorial construction of locally testable codes. In Richard E. Ladner and Cynthia Dwork, editors, *STOC*, pages 285–294. ACM, 2008.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006.

[RS96] Ronitt Rubinfeld and Madhu Sudan. Robust characterizations of polynomials with applications to program testing. *SIAM Journal on Computing*, 25(2):252–271, April 1996.