RC23465 (W0412-036) December 13, 2004
Mathematics

IBM Research Report

Non-Abelian Homomorphism Testing, and Distributions Close to Their Self-Convolutions

Michael Ben Or
School of Computer Science and Engineering
The Hebrew University
Jerusalem, 91904
Israel

Don Coppersmith
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Mike Luby
Digital Fountain
39141 Civic Center Drive
Suite 300
Fremont, CA 94538

Ronnitt Rubinfeld
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Non-Abelian Homomorphism Testing, and Distributions Close to their Self-Convolutions

Michael Ben Or[1] and Don Coppersmith[2] and Mike Luby[3] and Ronitt Rubinfeld[4]

[1] School of Computer Science and Engineering, The Hebrew University, Jerusalem, 91904, Israel. benor@cs.huji.ac.il
[2] IBM TJ Watson Research Center, Yorktown Heights, NY 10598, USA. dcopper@us.ibm.com, http://www.research.ibm.com/people/c/copper/
[3] Digital Fountain, 39141 Civic Center Dr., Ste. 300, Fremont, CA 94538, luby@digitalfountain.com
[4] MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA 02139, USA. ronitt@csail.mit.edu

Abstract. In this paper, we study two questions related to the problem of testing whether a function is close to a homomorphism. For two finite groups G, H (not necessarily Abelian), an arbitrary map $f : G \to H$, and a parameter $0 < \epsilon < 1$, say that f is ϵ-close to a homomorphism if there is some homomorphism g such that g and f differ on at most $\epsilon|G|$ elements of G, and say that f is ϵ-far otherwise. For a given f and ϵ, a homomorphism tester should distinguish whether f is a homomorphism, or if f is ϵ-far from a homomorphism. When G is Abelian, it was known that the test which picks $O(1/\epsilon)$ random pairs x, y and tests that $f(x) + f(y) = f(x + y)$ gives a homomorphism tester. Our first result shows that such a test works for all groups G.
Next, we consider functions that are close to their self-convolutions. Let $A = \{a_g | g \in G\}$ be a distribution on G. The self-convolution of A, $A' = \{a'_g | g \in G\}$, is defined by $a'_x = \sum_{y,z \in G; yz=x} a_y a_z$. It is known that $A = A'$ exactly when A is the uniform distribution over a subgroup of G. We show that there is a sense in which this characterization is robust – that is, if A is close in statistical distance to A', then A must be close to uniform over some subgroup of G. [1]

1 Introduction

In this paper, we focus on two questions that are related to the problem of testing whether a function is close to a homomorphism.

For two finite groups G, H (not necessarily Abelian), an arbitrary map $f : G \to H$, and a parameter $0 < \epsilon < 1$, say that f is ϵ-close to a homomorphism if there is some homomorphism g such that g and f differ on at most $\epsilon|G|$ elements of G. Define δ, the probability of group law failure, by

$$1 - \delta = \Pr_{x,y} [f(x) \times f(y) = f(x \times y)] .$$

Define τ such that τ is the minimum ϵ for which f is ϵ-close to a homomorphism. In [5], it was shown that over Abelian groups, there is a constant δ_0, such that if $\delta \leq \delta_0$, then the one can upper bound τ in terms of a function of δ that is independent of $|G|$. This yields a homomorphism tester with query complexity that depends (polynomially) on $1/\epsilon$, but is independent of $|G|$. In particular, the writeup in [5] contains an improved argument by Coppersmith [6], which shows that $\delta_0 < 2/9$ suffices, and that τ is upper bounded by the smaller root of $x(1 - x) = \delta$ (yielding a homomorphism tester with query complexity linear in $1/\epsilon$). Furthermore, the bound on δ_0 was shown to be tight for general groups [6].

Our first result is to give a relationship between the probability of group law failure and the closeness to being a homomorphism that applies to general (non-Abelian) groups. We show that for $\delta_0 < 2/9$, then f is τ-close to a homomorphism where $\tau = (3 - \sqrt{9 - 24\delta})/12 \leq \delta/2$ is the smaller root of $3x - 6x^2 = \delta$. The condition on δ, and the bound on τ as a function of δ, are shown to be tight, and the latter improves that of [5].

Next, consider the following question about distributions that are close to their self-convolutions: Let $A = \{a_g | g \in G\}$ be a distribution on group G. The convolution of distributions A, B is

$$C = A * B,\ c_x = \sum_{y,z \in G;\ yz=x} a_y b_z.$$

[1] A preliminary version of this paper appeared in RANDOM 2004 [3].

Let A' be the *self-convolution* of A, $A * A$, i.e. $a'_x = \sum_{y,z \in G; yz=x} a_y a_z$. It is known that $A = A'$ exactly when A is the uniform distribution over a subgroup of G. The question considered here is: when is A close to A'? In particular, if $dist(A, A') = \frac{1}{2}\sum_{x \in G} |a_x - a'_x| \leq \epsilon$ for small enough ϵ, what can be said about A? We show that A must be close to the uniform distribution over a subgroup of G, that is, for a distribution A over a group G, if $dist(A, A * A) \leq \epsilon \leq 0.0273$, then there is a subgroup H of G such that $dist(A, U_H) \leq 5\epsilon$, where U_H is the uniform distribution over H. On the other hand, we give an example of a distribution A such that $dist(A, A * A) \approx .1504$, but A is not close to uniform on any subgroup of the domain.

A weaker version of this result, with a somewhat more complicated proof, was used in the original proof of the homomorphism testing result in [5]. The earlier result was never published since the simpler and more efficient proof from [6] was substituted. Instead, a separate writeup of weaker versions of both of the results in this paper, by the current set of authors, was promised in [5]. This paper is the belated fulfillment of that promise, though the earlier results have been strengthened in the intervening time.

To give a hint of why one might consider the question on convolutions of distributions when investigating homomorphism testing, consider the distribution A_f achieved by picking x uniformly from G and outputting $f(x)$. It is easy to see that the error probability δ in the homomorphism test is at least $dist(A_f, A_f * A_f)$. Unfortunately, this last relationship is not in the useful direction. In Section 4, we present a relationship between homomorphism testing and distributions close to their self-convolution.

Related work: There has been interest in improving various parameters of homomorphism testing results, due to their use in the construction of Probabilistically Checkable Proof Systems (cf. [1]). In particular, both the constant δ_0 and the number of random bits required by the homomorphism test affect the efficiency of the proof system and in turn the hardness of approximation results that one can achieve using the proof system.

The homomorphism testing results can be improved in some cases: We have mentioned that $\delta_0 < 2/9$ is optimal over general Abelian groups [6]. However, using Fourier techniques, Bellare et. al. [2] have shown that for groups of the form $(\mathbf{Z}/2)^n$, $\delta_0 \leq 45/128$ suffices.

Several works have shown methods of reducing the number of random bits required by the homomorphism tests. That is, in the natural implementation of the homomorphism test, $2\log|G|$ random bits per trial are used to pick x, y. The results of [8, 7, 4, 9] have shown that fewer random bits are sufficient for implementing the homomorphism tests. The recent work of [9] gives a homomorphism test for general (non-Abelian) groups that uses only $(1 + o(1)) \log_2 |G|$ random bits. Given a Cayley graph that is an expander with normalized second eigenvalue γ, and for the analogous definitions of δ, τ, they show that for $\delta < (1 - \gamma)/12$, τ is upper bounded by $4\delta/(1 - \gamma)$.

2 Non-Abelian homomorphism testing

In this section, we show that the homomorphism test of [5] works over non-Abelian groups as well. As in the Introduction, we define δ, the probability of group law failure, by

$$1 - \delta = \Pr_{x,y}\left[f(x) \times f(y) = f(x \times y)\right].$$

We prove the following:

Theorem 1. *If $\delta < 2/9$ then f is τ-close to a homomorphism, where $\tau = [3 - \sqrt{9 - 24\delta}]/12 < \delta/2$ is the smaller root of $3x - 6x^2 = \delta$.*

The rest of this section is devoted to proving the theorem, and showing that the parameters of the theorem are tight.

$\Pr_{x,y}(\star)$ is the probability of $\star$ when x, y are independently selected from G with the uniform random distribution.

Given two finite groups G, H (not necessarily Abelian), and given an arbitrary map $f : G \to H$ (not necessarily a homomorphism) we will construct a map $g : G \to H$, which (under certain conditions on f) will be a group homomorphism and will agree with f on a large fraction of its domain G.

Given $f : G \to H$, with associated $\delta < 2/9$, we define $g : G \to H$ by

$$g(a) = majority_{x \in G}\left[f(a \times x) \times f(x)^{-1}\right].$$

That is, we evaluate the bracketed expression for each $x \in G$, and let $g(a)$ be the value most often attained. Define ϵ_a, ϵ and τ:

$$1 - \epsilon_a = \Pr_x \left[f(a \times x) f(x)^{-1} = g(a) \right]$$
$$\epsilon = \max_a \epsilon_a$$
$$1 - \tau = \Pr_x \left[f(x) = g(x) \right]$$

Lemma 1. *If $\delta < 2/9$ then $\epsilon_a \leq \hat{\epsilon}$ where $\hat{\epsilon}$ is the smaller root of $x - x^2 = \delta$.*

Proof: For $a \in G$, define

$$p_a = \Pr_{x,y \in G} \left[f(a \times x) \times f(x)^{-1} = f(a \times y) \times f(y)^{-1} \right].$$

By rearranging, we have

$$\begin{aligned} p_a &= \Pr_{x,y \in G} \left[f(a \times y)^{-1} \times f(a \times x) = f(y)^{-1} \times f(x) \right] \\ &\geq \Pr_{x,y \in G} \left[f(a \times y)^{-1} \times f(a \times x) = f(y^{-1} \times x) \wedge f(y)^{-1} \times f(x) = f(y^{-1} \times x) \right]. \end{aligned}$$

Each of the latter two equations is a random instance of the test equation $f(u) \times f(v) \stackrel{?}{=} f(u \times v)$, or equivalently, $f(u)^{-1} \times f(u \times v) \stackrel{?}{=} f(v)$, so each holds with probability $1-\delta$, and, by the union bound, they both hold simultaneously with probability at least $1 - 2\delta$. So we have

$$p_a \geq 1 - 2\delta > 5/9.$$

If we partition G into blocks

$$B_{a,z} = \{x \in G | f(a \times x) \times f(x)^{-1} = z\}$$

with relative sizes $b_{a,z} = |B_{a,z}|/|G|$, then

$$\sum_z b_{a,z} = 1$$
$$p_a = \sum_z b_{a,z}^2 \leq \max_z b_{a,z}$$

so that $\max_z(b_{a,z}) > 5/9$, and $g(a) = argmax_z(b_{a,z})$ is well defined. By definition, $1 - \epsilon_a = \max_z(b_{a,z}) > 5/9$. Since $1 - \epsilon_a > 1/2$, we also have

$$p_a \leq (1 - \epsilon_a)^2 + \epsilon_a^2$$
$$1 - 2\delta \leq 1 - 2\epsilon_a + 2\epsilon_a^2$$
$$\delta \geq \epsilon_a - \epsilon_a^2,$$

and since $\epsilon_a < 1/2$, we conclude that $\epsilon_a \leq \hat{\epsilon}$, the smaller root of $x - x^2 = \delta$, as required. □

Corollary 1. *If $\delta < 2/9$ then $\epsilon_a < 1/3$ and $\epsilon < 1/3$.*

Lemma 2. *If $\delta < 2/9$ then g is a homomorphism.*

Proof:

IDENTITY: $g(1) = 1$. Immediate since each value x gives $f(1 \times x) \times f(x)^{-1} = 1$.

INVERSE: $g(a^{-1}) = g(a)^{-1}$. There is a one-one correspondence between x satisfying $f(a \times x) \times f(x)^{-1} = g(a)$ and y satisfying $f(a^{-1} \times y) \times f(y)^{-1} = g(a)^{-1}$, namely $y = a \times x$.

PRODUCT: $g(a) \times g(b) = g(a \times b)$. Each of the following three equations holds with probability at least $1 - \epsilon > 2/3$ on random choice of y:

$$g(a) = f(a \times y) \times f(y)^{-1}$$
$$g(b) = f(y) \times f(b^{-1} \times y)^{-1}$$
$$g(a \times b) = f(a \times y) \times f(b^{-1} \times y)^{-1}$$

(In the definition of g, we substitute $y = x$ in the first equation, and $y = b \times x$ in the second and third.) By the union bound, all three equations hold simultaneously with probability at least $1 - 3\epsilon > 0$; that is, there is at least one value of y satisfying all three equations. Substituting one such value of y and combining these three equations, we conclude $g(a) \times g(b) = g(a \times b)$, as desired. □

Lemma 3. $\tau \leq \delta + \epsilon$.

Proof:

$$\begin{aligned}\tau &= \Pr_a [f(a) \neq g(a)] \\ &\leq \Pr_{a,x} \left[f(a) \neq f(a \times x) \times f(x)^{-1}\right] + \Pr_{a,x} \left[g(a) \neq f(a \times x) \times f(x)^{-1}\right] \\ &\leq \delta + Average_a(\epsilon_a) \leq \delta + \epsilon.\end{aligned}$$

□

Lemma 4. $\epsilon \geq 2(\tau - \tau^2)$.

Proof:

$$\Pr_{x,y} \left[f(x) \times f(y)^{-1} \neq g(x \times y^{-1})\right]$$

is the average value of ϵ_a (over random choices of a), and so is bounded by ϵ. This group law failure will hold at least if either of these two mutually exclusive events occurs, since g is a homomorphism:

- $f(x) = g(x) \wedge f(y) \neq g(y)$;
- $f(x) \neq g(x) \wedge f(y) = g(y)$.

By the independence of x, y, each of the two events has probability $\tau(1 - \tau)$. So

$$\epsilon \geq \tau(1 - \tau) + (1 - \tau)\tau = 2(\tau - \tau^2).$$

□

Corollary 2. *If* $\delta < 2/9$ *then* $\tau < \frac{3-\sqrt{3}}{6} < 0.2114$.

Proof: If $\delta < 2/9$ then $\epsilon < 1/3$, and Lemma 4 implies either $\tau < \frac{3-\sqrt{3}}{6} < 0.2114$ or $\tau > \frac{3+\sqrt{3}}{6} > 0.7886$. The latter is inconsistent with $\tau \leq \delta + \epsilon < 5/9$ (Lemma 3). □

Lemma 5. *If* $\delta < 2/9$ *then* $\delta \geq 3\tau - 6\tau^2$.

Proof: Since g is a homomorphism, the inequality $f(x) \times f(y) \neq f(x \times y)$ (which has probability δ) will hold in at least the following three mutually exclusive events:

- $f(x) = g(x) \wedge f(y) = g(y) \wedge f(x \times y) \neq g(x \times y)$;
- $f(x) = g(x) \wedge f(y) \neq g(y) \wedge f(x \times y) = g(x \times y)$;
- $f(x) \neq g(x) \wedge f(y) = g(y) \wedge f(x \times y) = g(x \times y)$.

Each event has probability at least $\tau - 2\tau^2$. Taking the first for example, we have (by pairwise independence of the arguments)

$$\Pr_{x,y} [f(x \times y) \neq g(x \times y)] = \tau$$

$$\Pr_{x,y} [f(x \times y) \neq g(x \times y) \wedge f(x) \neq g(x)] = \tau^2$$

$$\Pr_{x,y} [f(x \times y) \neq g(x \times y) \wedge f(y) \neq g(y)] = \tau^2$$

$$\Pr_{x,y} [f(x \times y) \neq g(x \times y) \wedge \{f(x) \neq g(x) \vee f(y) \neq g(y)\}] \leq 2\tau^2$$

$$\Pr_{x,y} [f(x \times y) \neq g(x \times y) \wedge \{f(x) = g(x) \wedge f(y) = g(y)\}] \geq \tau - 2\tau^2$$

Since the three events are mutually exclusive, their probabilities add, giving

$$\delta \geq 3(\tau - 2\tau^2).$$

□

Lemma 6. *If* $\delta < 2/9$ *then* τ *is bounded by the smaller root of* $3x - 6x^2 = \delta$.

Proof: Combine Corollary 2 ($\tau < 0.2114$) with Lemma 5 ($\delta \geq 3\tau - 6\tau^2$). □

This finishes the proof of Theorem 1.

We have the following complement to Lemma 5:

Lemma 7. *If $\delta < 2/9$ then $\delta \leq 3\tau - 2\tau^2$.*

Proof: For random x, y, the three events $f(x) \neq g(x)$, $f(y) \neq g(y)$ and $f(xy) \neq g(xy)$ are pairwise independent, and each has probability τ. If $f(x)f(y) \neq f(xy)$ then at least one of the following holds:

- $f(x) \neq g(x)$;
- $f(y) \neq g(y) \wedge f(x) = g(x)$;
- $f(xy) \neq g(xy) \wedge f(x) = g(x)$.

Summing the probabilities of these three events,

$$\delta = \Pr_{x,y}[f(xy) \neq f(x)f(y)] \leq \tau + (\tau - \tau^2) + (\tau - \tau^2) = 3\tau - 2\tau^2.$$

□

This gives us a tight estimate of τ in terms of δ:

Corollary 3. *If $\delta < 2/9$ then $\tau = \delta/3 + O(\delta^2)$.*

Proof: Combine Lemmas 6 and 7. □

Example 1. The bound $\delta < 2/9$ is tight. The following example has $\delta = 2/9$ and $\epsilon = 1/3$, but $\tau = 1 - 1/3^{k-1}$ is arbitrarily close to 0. Here the groups are written additively rather than multiplicatively.

$$f : \mathbf{Z}/3^k \to \mathbf{Z}/3^{k-1}$$

$$f(3\ell + d) = \ell, 0 \leq \ell < 3^{k-1}, d \in \{-1, 0, 1\}$$

The group law $f(x) + f(y) = f(x + y)$ is violated exactly when $x = 3\ell + d$, $y = 3m + d$, $x + y = 3(\ell + m + d) - d \bmod 3^k$, $d \in \{-1, 1\}$, which happens with probability exactly 2/9. Each homomorphism $g_j : \mathbf{Z}/3^k \to \mathbf{Z}/3^{k-1}$ is given by an integer $j \in \{0, 1, \ldots, 3^{k-1} - 1\}$, namely $g_j(m) = jm \pmod{3^{k-1}}, 0 \leq m < 3^k$. Each homomorphism g_j agrees with f in exactly three arguments: if $f(3m + d) = g_j(3m + d)$ with $d \in \{-1, 0, 1\}$ and $0 \leq m < 3^{k-1}$, then

$$j(3m + d) = m \pmod{3^{k-1}}$$

$$m = dj/(1 - 3j) \pmod{3^{k-1}},$$

since $1 - 3j$ is invertible $\pmod{3^{k-1}}$; so that for each of three possible values of d we have exactly one argument $3m + d$ where the two maps agree. This yields $\tau = 1 - 3/3^k$.

Example 2. The bound $3\tau - 6\tau^2 \leq \delta$ is tight. Choose τ' with $0 < \tau' \leq 1/3$, choose N an arbitrarily large odd positive integer, and define $f : \mathbf{Z}/N \to \mathbf{Z}/2$ by

$$f(x) = 1 \Leftrightarrow \tau' < \frac{x}{N} < 2\tau'.$$

(Again the groups are written additively.) Since N is odd, the only possible homomorphism g is $g(x) \equiv 0$. We have

$$\tau = \Pr_x[f(x) = 1] = \tau' + O(1/N).$$

The error $O(1/N)$ is due to rounding errors.

$$\Pr_{x,y}[f(x) = f(y) = 1] = \tau^2.$$

This comes from independence of x, y.

$$\Pr_{x,y}[f(x) = f(y) = f(x + y) = 1] = 0.$$

The third equation comes from the fact that $x/N, y/N, (x+y)/N$ cannot simultaneously lie in the interval $(\tau', 2\tau')$, nor can $x/N, y/N, (x+y-N)/N$. So an error $f(x)+f(y) \neq f(x+y)$ will happen precisely when exactly one of $f(x), f(y), f(x+y)$ is 1. We calculate from the above equations that

$$\Pr_{x,y}[f(x)=1, f(y)=0, f(x+y)=0] = \tau - 2\tau^2;$$

similarly,

$$\Pr_{x,y}[f(x)=0, f(y)=1, f(x+y)=0] = \tau - 2\tau^2,$$

$$\Pr_{x,y}[f(x)=0, f(y)=0, f(x+y)=1] = \tau - 2\tau^2.$$

These are the only three ways the group law failure can arise, and they are mutually exclusive, yielding:

$$\delta = 3(\tau - 2\tau^2).$$

3 Convolutions of distributions

In this section, we show that for a distribution A over a finite group G, if $|A - A * A| \leq \epsilon$ then A is δ-close to the uniform distribution over a subgroup of G.

We let capital letters A, B, C denote distributions over group G and subscripted uncapitalized letters a_x, b_y denote the probability of a particular element. X, Y, Z, H will be subsets of G.

We let U_S denote the uniform distribution on $S \subseteq G$.

We let $dist(A,B) = \frac{1}{2}|A-B|$. Note that distances satisfy the triangle inequality, i.e., $dist(A,C) \leq dist(A,B) + dist(B,C)$. Also it is easy to see that $dist(A*B, A*C) \leq dist(B,C)$.

It will also be convenient to consider a second kind of convolution,

$$C = A \bullet B,\ c_x = \sum_{y,z \in G;\ xy=z} a_y b_z.$$

When we have uniform distributions on subsets of equal size, the two convolutions enjoy the following relation:

Lemma 8. *Let X, Y, Z be subsets of a finite group G, with $|X| = |Y| = |Z| = n$. Then*

$$dist(U_X, U_Y * U_Z) = dist(U_Y, U_Z \bullet U_X).$$

Proof: For any two distributions A, B, since $|A| = |B| = 1$, we have

$$dist(A,B) = \sum_{a_x > b_x} (a_x - b_x) = \sum_{a_x < b_x} (b_x - a_x).$$

Then

$$\begin{aligned}
dist(U_X, U_Y * U_Z) &= \textstyle\sum_{x \in X}[(U_X)_x - (U_Y * U_Z)_x] \\
&= \textstyle\sum_{x \in X}[\frac{1}{n} - \sum_{y \in Y, z \in Z, yz=x} \left(\frac{1}{n}\right)\left(\frac{1}{n}\right)] \\
&= 1 - \frac{1}{n^2}|\{(x,y,z) : x \in X, y \in Y, z \in Z, yz = x\}| \\
&= \textstyle\sum_{y \in Y}[\frac{1}{n} - \sum_{z \in Z, x \in X, yz=x} \left(\frac{1}{n}\right)\left(\frac{1}{n}\right)] \\
&= \textstyle\sum_{y \in Y}[(U_Y)_y - (U_Z \bullet U_X)_y] \\
&= dist(U_Y, U_Z \bullet U_X).
\end{aligned}$$

□

Remark 1. The lemma does not hold for arbitrary distributions, nor for uniform distributions on subsets of different sizes.

Overview of proof: We will embed G in a larger group $F = G \times \mathbf{Z}/N$ for suitably large N, and consider a distribution B induced from A, namely $b_{(x,j)} = a_x/N$. This will alleviate problems later when we have to round to integers. We show that if $B' = B * B$ is close to B, then there is a set $X \subseteq F$ such that B is close to U_X. We next show that X must be close to a subgroup $\hat{H}$ of F, and further that this subgroup is of the form $\hat{H} = H \times \mathbf{Z}/N$. Then B is close to the uniform distribution on $\hat{H}$, and A is close to the uniform distribution on H. A bootstrap lemma allows us to claim that once A is moderately close to U_H, then it is very close.

Expanding the group: Pick N suitably large. Define $F = G \times \mathbf{Z}/N$, with elements $\{(x,j) : x \in G, j \in \mathbf{Z}/N\}$ and group law $(x,j)(y,k) = (xy, j+k)$. The distribution B on F is given by $A \times U_{\mathbf{Z}/n}$, that is, $b_{(x,j)} = a_x(1/N)$. Defining $B' = B * B$ and $A' = A * A$, it is immediate that $dist(B,B') = dist(A,A')$.

B is close to uniform on a subset: Our first theorem shows that if $B' = B * B$ is close to B, then there is a set $X \subseteq F$ such that B is close to U_X.

Theorem 2. *Let F be a finite group. Let B be a distribution on F for which no element has probability more than $1/N$. Let $1/8 > \epsilon > 0$ be a constant. If $dist(B, B*B) \leq \epsilon$ then there is a set $X \subseteq G$ such that $dist(B, U_X) \leq \epsilon'$ where $\epsilon' = 3\epsilon + O(1/N)$. Further, $dist(U_X, U_X * U_X) \leq 6\epsilon + O(\epsilon^2) + O(1/N)$.*

Proof: Let $B' = B * B$. In the rest of the proof, relabel the elements such that $b_1 \geq b_2 \geq b_3 \geq \ldots$, i.e., 1 corresponds to the element of F with the highest probability mass. For given $x \in G$, the N elements $b_{(x,k)}, k \in \mathbf{Z}/N$, are equal, so we arrange that they are contiguous in this ordering.

For $n \geq 1$, let $\mathtt{sum}_n = \sum_{j=1}^{n} b_j$ be the sum of the n highest probabilities. Also, let $\mathtt{sum}'_n = \sum_{j=1}^{n} b'_j$ be the sum of the probabilities with respect to B' of the n most likely elements with respect to B.

Let $\alpha_n = \mathtt{sum}_n^2 + n\sum_{j>n} b_j^2$. It is not hard to see that $\mathtt{sum}_n \geq \alpha_n$.

Claim. $\alpha_n \geq \mathtt{sum}'_n$.

Proof: [of claim] Construct a bipartite graph, where $U = \{u_s : s \in F\}$ and $V = \{v_s : s \in F\}$ are disjoint sets of nodes. For all $s \in F$, let $E_s = \{(u_x, v_y) : x * y = s\}$. Let the weight of each edge (u_x, u_y) in E_s be $b_x b_y$, and the weight of E_s be $wt(E_s) = b'_s = \sum_{(u_x,v_y)\in E_s} b_x b_y$. Let $\mathcal{G} = (U, V, E)$ where $E = \bigcup_{s\in F} E_s$. Since F is a group, each vertex is of degree $|F|$, and there are no multiple edges. The weight of $\mathcal{G}$ is the total weight of all edges, i.e., $wt(\mathcal{G}_F) = \sum_s b'_s = 1$.

Because of the initial relabeling of B, u_j and v_j correspond to the elements with the j^{th} highest probability mass. Let $\mathcal{G}_n$ be the graph induced on vertices $\{1 \ldots n\}$ according to this relabeling. Notice that $wt(\mathcal{G}_n) = \mathtt{sum}'_n$.

We show how to transform $\mathcal{G}_n$ into the graph $\mathcal{G}'_n$ by a series of edge swaps such that $wt(\mathcal{G}_n) \leq wt(\mathcal{G}'_n)$, where $\mathcal{G}'_n$ is the complete bipartite graph between vertices $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$, along with n multiple edges between u_j and v_j for each $j > n$. From this the claim follows since $\alpha_n = wt(\mathcal{G}'_n)$.

> **Edge Swap:** Let $k, k', l, l' \geq 1$ be such that $k < l', k' < l$. Suppose the edges (u_k, v_l) and $(u_{l'}, v_{k'})$ exist. The swap consists of deleting these two edges and adding the two new edges $(u_k, v_{k'})$ and $(u_{l'}, v_l)$.

After each edge swap, the new weight minus the old weight is $(b_k - b_{l'})(b_{k'} - b_l)$. This is nonnegative because $k < l'$ implies that $b_k \geq b_{l'}$ and $k' < l$ implies that $b_{k'} \geq b_l$.

The first part of the swap sequence to go from $\mathcal{G}_n$ to $\mathcal{G}'_n$ is as follows. For all $k, k' \leq n$ such that there is no edge $(u_k, v_{k'})$, there must exist an $l > n$ and an $l' > n$ such that edges (u_k, v_l) and $(u_{l'}, v_{k'})$ exist. This is because each vertex has degree n in $\mathcal{G}_n$ and initially there are no multiple edges among the first n vertices and we retain these properties throughout the swap sequence. Use an edge swap to delete these two edges and add the two edges $(u_k, v_{k'})$ and $(u_{l'}, v_l)$. Note that $(u_k, v_{k'})$ is not a multiple edge, although $(u_{l'}, v_l)$ might be. Still, since $l, l' > n$ the swap does not create multiple edges among the first n vertices on each side of the bipartition. This sequence of swaps creates a complete bipartite graph among the first n vertices on each side of the bipartition.

The rest of the swap sequence is as follows. For all $j > n$, if there are not n multiple edges from u_j to v_j then there must be an $l > j$ and an $l' > j$ such that (u_j, v_l) and $(u_{l'}, v_j)$ are both edges. Use an edge swap to delete these two edges and add the edges (u_j, v_j) and $(u_{l'}, v_l)$. This eventually reaches $\mathcal{G}'$, thus proving the claim. □

We will define an X such that $dist(B, U_X)$ is small. Pick τ with $1/4 \leq \tau \leq 3/4$; later we will specify $\tau = 3/5$. Select m with $\mathtt{sum}_{m-1} < \tau \leq \mathtt{sum}_m$. Set $h = b_m$. Set $p = \lfloor 1/h \rfloor$. Let the distribution $\hat{U}$ assign weight h to the first p elements, and a weight $1 - ph < h$ to the $(p+1)$st element. Let $n = p$ if $b_{p+1} > \hat{u}_{p+1}$, and $n = p+1$ otherwise. Let X consist of the first n elements, so that $dist(\hat{U}, U_X) < h = O(1/N)$. Also define $g = b_n$.

The distribution B differs from $\hat{U}$ in three places. For $i \leq m$, $b_i \geq \hat{u}_i$; define $\beta = \sum_{i\leq m}(b_i - \hat{u}_i)$. For $i > n$, $b_i \geq \hat{u}_i$; define $\delta = \sum_{i>n}(b_i - \hat{u}_i)$. For $m < i \leq n$, $b_i \leq \hat{u}_i$; we have $\beta + \delta = \sum_{m<i\leq n}(\hat{u}_i - b_i)$. So $dist(B, \hat{U}) = \beta + \delta$.

For $m < i \leq n$ we have $g \leq b_i \leq h$, so that $b_i^2 \leq (g+h)b_i - gh$. Similarly for $i > n$ we have $0 \leq b_i \leq g$, so that $b_i^2 \leq g b_i$. Recall also that $mh + \beta = \tau + O(1/N)$ and $nh = 1 + O(1/N)$. This enables the following computation:

$$
\begin{aligned}
\epsilon &\geq \mathtt{sum}_m - \mathtt{sum}'_m \\
&\geq \mathtt{sum}_m - \alpha_m \\
&= (mh+\beta) - (mh+\beta)^2 - m\textstyle\sum_{m+1}^{n} b_i^2 - m\sum_{i>n} b_i^2 \\
&\geq (mh+\beta)(1-mh-\beta) - m\textstyle\sum_{m+1}^{n}[(g+h)b_i - gh] - m\sum_{i>n} gb_i \\
&= (mh+\beta)(1-mh-\beta) - m(g+h)\textstyle\sum_{m+1}^{n} b_i + m(n-m)gh - mg\sum_{i>n} b_i \\
&= (mh+\beta)(1-mh-\beta) - m(g+h)(nh-mh-\beta-\delta) + mngh - m^2gh - mg\delta \\
&= (mh+\beta)(1-mh-\beta) - m(g+h)(1-mh-\beta-\delta) + mg - m^2gh - mg\delta + O(1/N) \\
&= (1-mh-\beta)(mh+\beta-mg-mh) + m(g+h)(\delta) + mg - m^2gh - mg\delta + O(1/N) \\
&= (1-mh-\beta)(\beta) - (1-mh-\beta)(mg) + mh\delta + mg - m^2gh + O(1/N) \\
&= (1-\tau)(\beta) + mh\delta + mg\beta + O(1/N) \\
&= (1-\tau)(\beta) + (\tau-\beta)\delta + mg\beta + O(1/N) \\
&\geq (1-\tau)\beta + \tau\delta - \beta\delta + O(1/N) \\
&= [(1-\tau)\beta + \tau\delta] + \tfrac{1}{4\tau(1-\tau)}\{[(1-\tau)\beta - \tau\delta]^2 - [(1-\tau)\beta + \tau\delta]^2\} + O(1/N) \\
&\geq [(1-\tau)\beta + \tau\delta] - \tfrac{1}{4\tau(1-\tau)}[(1-\tau)\beta + \tau\delta]^2 + O(1/N) \\
&= u - \tfrac{u^2}{4\tau(1-\tau)} + O(1/N)
\end{aligned}
$$

where $u = (1-\tau)\beta + \tau\delta$. Now $\beta \leq \tau + O(1/N)$ and $\delta \leq 1-\tau$ so that $u = (1-\tau)\beta + \tau\delta \leq 2\tau(1-\tau)$ is less than the larger root of

$$\frac{x^2}{4\tau(1-\tau)} - x + \epsilon = 0,$$

namely

$$[(1-\tau)\beta + \tau\delta] \leq 2\tau(1-\tau) + O(1/N) \leq 2\tau(1-\tau)[1 + \sqrt{1 - \epsilon/\tau(1-\tau)}]$$

(since $\epsilon/\tau(1-\tau) < (1/8)/(1/4 \times 3/4) < 1$), so it must be less than the smaller root:

$$[(1-\tau)\beta + \tau\delta] \leq 2\tau(1-\tau)[1 - \sqrt{1 - \epsilon/\tau(1-\tau)}] + O(1/N),$$

remembering the error term.

Substituting $\tau = 3/5$, we have

$$2\beta + 3\delta \leq \frac{12}{5}\left[1 - \sqrt{1 - \frac{25}{6}\epsilon}\right] + O(1/N) = 5\epsilon + O(\epsilon^2) + O(1/N).$$

By the triangle inequality,

$$dist(B, U_X) \leq dist(B, \hat{U}) + dist(\hat{U}, U_X) = \beta + \delta + O(1/N).$$

For small ϵ we can estimate the error: If $\epsilon < 1/8$ then

$$\beta + \delta \leq \frac{1}{2}(2\beta + 3\delta) \leq \frac{1}{2} \cdot \frac{12}{5}\left[1 - \sqrt{1 - \frac{25}{6}\epsilon}\right] + O(1/N) < 3\epsilon + O(1/N),$$

establishing the theorem. For later use, we also note that if $\epsilon < 0.0273$ then $\beta + \delta < 2.6\epsilon$.

Repeatedly applying the triangle inequality, we can obtain:

$$dist(B, U_X * U_X) \leq dist(B, B * B) + dist(B, B * U_X) + dist(B, U_X * U_X) \leq \epsilon + 2(\beta + \delta).$$

To obtain $dist(U_X, U_X * U_X)$ we could apply the triangle inequality again. Instead, we recall that B differs from $\hat{U}$, and hence U_X, in three pieces, namely (up to errors of order $O(1/N)$) β (before m), $-\beta - \delta$ (between m and n), and δ (after n). The convolution $U_X * U_X$ is everywhere bounded by $n(1/n)^2 = 1/n$. So when we change B to U_X and monitor the change in $dist(\star, U_X * U_X)$, the first change of β is driving us closer to $U_X * U_X$, while the other two changes of $\beta + \delta$ and δ might drive us further away. The net result is

$$
\begin{aligned}
dist(U_X, U_X * U_X) &\leq dist(B, U_X * U_X) + \tfrac{1}{2}\left[-\beta + (\beta + \delta) + \delta\right] + O(1/N) \\
&\leq \epsilon + 2\beta + 3\delta + O(1/N) \\
&\leq \epsilon + \tfrac{12}{5}\left[1 - \sqrt{1 - \tfrac{25}{6}\epsilon}\right] + O(1/N) \\
&= 6\epsilon + O(\epsilon^2) + O(1/N).
\end{aligned}
$$

This establishes the second part of the theorem.□

B is close to uniform on a subgroup of F: Next we show that if the uniform distribution on X is close in distance to its convolution with itself, then X is close to some subgroup $\hat{H}$ of F.

Theorem 3. *Let F be a finite group and X a subset of F. Let $\tau = dist(U_X, U_X * U_X) = dist(U_X, U_X \bullet U_X)$. If $\tau < 1/9$ then there is a subgroup $\hat{H}$ of F with $|X \backslash \hat{H}| + |\hat{H} \backslash X| \leq 3\tau|X|$.*

Proof: Let $n = |X|$. Let $V = U_X \bullet U_X$ so that

$$v_x = \frac{1}{n^2}|\{(y,z) : y, z \in X, xy = z\}|.$$

If e is the identity element of F, we see $v_e = \frac{1}{n}$, and $v_x \leq \frac{1}{n}$ for all $x \in G$.

We need to establish a triangle inequality on quantities such as $(v_e - v_x)$.

Lemma 9. *For $x, y \in F$, the quantities $(v_e - v_x)$, $(v_e - v_y)$, $(v_e - v_{xy})$ are nonnegative and satisfy the triangle inequalities:*

$$(v_e - v_x) + (v_e - v_y) \geq (v_e - v_{xy}) \tag{1}$$

$$(v_e - v_x) + (v_e - v_{xy}) \geq (v_e - v_y) \tag{2}$$

$$(v_e - v_y) + (v_e - v_{xy}) \geq (v_e - v_x) \tag{3}$$

Proof: If $v_x = k_x/n^2$, with $0 \leq k_x \leq n$, then there are k_x elements $z \in F$ such that both z and xz are in X; call such z "good elements" for x. There are $n - k_x$ elements z such that $z \in X$ and $xz \notin X$; there are $n - k_x$ elements z such that $z \notin X$ and $xz \in X$; call the latter two kinds of z "bad elements" for x. The number of bad elements for x is $2(n - k_x) = 2n^2(v_e - v_x)$. If z is neither good nor bad for x it is "neutral" for x.

For each $z \in F$, consider the three elements z, yz, xyz. If all three are in X, then we have found good elements for each of x, y, xy. (Namely, z is a good element for y and for xy, and yz is a good element for x.) If exactly two are in X, then we have found bad elements for exactly two of x, y, xy, and a good element for the other. (For example, if $z, xyz \in X$ and $yz \notin X$, then z is good for xy, z is bad for y, and yz is bad for x.) If exactly one of z, yz, xyz is in X, then we have bad elements for exactly two of x, y, xy and a neutral one for the other. If $z, yz, xyz \notin X$ then we have found neutral elements for all three. The important point is that the "bad elements" come in pairs, each time contributing to two of $2(n - k_x), 2(n - k_y), 2(n - k_{xy})$. Setting

$$\begin{aligned} p &= |\{z : z \text{ bad for } xy \text{ and } y\}| \\ q &= |\{z : z \text{ bad for } xy; yz \text{ bad for } x\}| \\ r &= |\{z : z \text{ bad for } y; yz \text{ bad for } x\}|, \end{aligned}$$

we find

$$\begin{aligned} 2(n - k_x) &= q + r \\ 2(n - k_y) &= p + r \\ 2(n - k_{xy}) &= p + q. \end{aligned}$$

This establishes the triangle inequality among $2(n - k_x), 2(n - k_y), 2(n - k_{xy})$, and hence the lemma. □

Next we show an "excluded middle" result for V.

Lemma 10. *Let $\tau < 1/6$. For all $x \in F$, either $v_x < \frac{1}{3n}$ or $v_x > \frac{2}{3n}$.*

Proof: Fix x in F. There are $n^2 \cdot v_x$ pairs (y, z) with $y, z \in X$ and $xy = z$. For each such pair, $(1/n - v_y) + (1/n - v_z) \geq (1/n - v_x)$, so $((U_X)_y - v_y) + ((U_X)_z - v_z) \geq 1/n - v_x$. There are $n - n^2 \cdot v_x = n(1 - nv_x)$ pairs (y, z) with $y \in X$, and $z \notin X$, and $xy = z$. For each such pair, $(1/n - v_y) + (1/n - v_x) \geq (1/n - v_z)$, so $(1/n - v_y) + v_z \geq v_x$, and $|(U_X)_y - v_y| + |(U_X)_z - v_z| \geq v_x$. There are $n(1 - nv_x)$ pairs (y, z) with $y \notin X$, and $z \in X$, and $xy = z$; again $|(U_X)_y - v_y| + |(U_X)_z - v_z| \geq v_x$. Summing all three cases, the left-hand side hits each element of X twice (once as y, once as z), and each element of $F - X$ at most twice, and we get $2\tau + 2\tau \geq n^2(v_x)(1/n - v_x) + 2n(1 - nv_x)(v_x)$. So $4\tau \geq 3(nv_x)(1 - (nv_x))$. If $\tau < 1/6$, this implies (nv_x) is either $< 1/3$ or $> 2/3$. □

The excluded middle gives us a natural candidate for our subgroup $\hat{H}$.

Lemma 11. *Let $\tau < 1/6$. Define $\hat{H} = \{x \in F : v_x > \frac{2}{3n}\}$. Then $\hat{H}$ is a subgroup of F.*

Proof: $\hat{H}$ contains the identity e because $v_e = 1/n$. $\hat{H}$ is closed under inverse: z is good for x if and only if xz is good for x^{-1}, whence $k_x = k_{x^{-1}}$ and $v_x = v_{x^{-1}}$, so $x \in H \Leftrightarrow x^{-1} \in H$. Closure under the group operation follows from the triangle inequality (1): if $x, y \in \hat{H}$ then

$$v_{xy} \geq v_x + v_y - v_e > \frac{2}{3n} + \frac{2}{3n} - \frac{1}{n} = \frac{1}{3n}$$

so that (by the excluded middle) $v_{xy} > \frac{2}{3n}$ and $xy \in \hat{H}$. □

Finally we show that $\hat{H}$ is close to X.

Lemma 12. *With $\tau < 1/6$ and $\hat{H}$ as above, the symmetric difference between X and $\hat{H}$ satisfies:*

$$|\hat{H} \backslash X| + |X \backslash \hat{H}| \leq \frac{2\tau}{1-3\tau}|X| \leq 3\tau|X|.$$

Proof: We have

$$\begin{aligned} 2\tau = 2dist(U_X, V) &= \textstyle\sum_{x \in F} |(U_X)_x - v_x| \\ &= \textstyle\sum_{x \in X}(\frac{1}{n} - v_x) + \sum_{x \in F \backslash X}(v_x - 0) \\ &\geq \textstyle\sum_{x \in X \backslash \hat{H}}(\frac{1}{n} - \frac{3\tau}{n}) + \sum_{x \in \hat{H} \backslash X}(\frac{1}{n} - \frac{3\tau}{n}) \\ &= \frac{1-3\tau}{n}\left[|X \backslash \hat{H}| + |\hat{H} \backslash X|\right] \end{aligned}$$

which, with $|X| = n$, proves the lemma. □

This establishes the theorem.□

B is close to a subgroup $\hat{H}$: We can push these results back to the distributions B:

Theorem 4. *Let F be a finite group. Let B be a distribution on F. Let $X \subseteq F$ be a subset. Given ϵ, β, δ such that:*

- $dist(B, B * B) \leq \epsilon$;
- $dist(B, U_X) \leq \beta + \delta$ *(as in Theorem 2);*
- $\tau = \epsilon + 2\beta + 3\delta < 1/6$,

then there is a subgroup $\hat{H}$ of F with $dist(B, U_{\hat{H}}) < \beta + \delta + 2\tau/(1-3\tau)$. The same is true if we replace the first condition with

$$dist(B, B \bullet B) \leq \epsilon.$$

Proof: As in Theorem 2, we have we have:

$$dist(U_X, U_X * U_X) \leq \epsilon + 2\beta + 3\delta + O(1/N) = \tau + O(1/N).$$

(This remains true if we replace "$*$" by "$\bullet$" throughout.) Then by Theorem 3, we construct a subgroup $\hat{H}$ with $|X \backslash \hat{H}| + |\hat{H} \backslash X| \leq \frac{2\tau}{1-3\tau}|X|$.

A direct calculation shows that if $|X| = n$, $|X \backslash \hat{H}| = b$, and $|\hat{H} \backslash X| = a$, then

$$dist(U_X, U_{\hat{H}}) = \begin{cases} a/(n+a-b) & \text{if } a \geq b \\ b/n & \text{if } a \leq b \end{cases}$$

For a fixed value of $a + b$ this distance is maximized when $a = 0$; so we have

$$dist(U_X, U_{\hat{H}}) \leq \frac{1}{n} \times \frac{2\tau n}{1-3\tau} = \frac{2\tau}{1-3\tau}.$$

Finally, $dist(B, U_{\hat{H}}) \leq dist(B, U_X) + dist(U_X, U_{\hat{H}}) \leq \beta + \delta + \frac{2\tau}{1-3\tau}$. □

Reverting to original group: The group $\hat{H}$ respects the block structure of $F = G \times \mathbf{Z}/N$, in the sense that for all $x \neq e \in G$ and $j, k \in \mathbf{Z}/N$, $v_{(x,j)} = v_{(x,k)}$, with v as defined in Lemma 9, so that $(x, j) \in \hat{H} \Leftrightarrow (x, k) \in \hat{H}$. Further, one can verify that $v_{(e,k)} \geq v_{(x,k)}$, so that if any $(x, k) \in \hat{H}$ with $x \neq e$, then the entire block $\{(e, k) : k \in \mathbf{Z}/N\}$ is in $\hat{H}$. (Note that if it were the case that there were no $(x, k) \in \hat{H}$ with $x \neq e$, then $H = \{e\}$, which is a subgroup of G.) This implies that $\hat{H}$ is of the form

$$\hat{H} = H \times \mathbf{Z}/N.$$

It is obvious that H is a subgroup of G, and that

$$dist(A, U_H) = dist(B, U_{\hat{H}}).$$

We tie in with Theorem 2.

Theorem 5. *Let A be a distribution on the finite group G. Let $dist(A, A * A) \leq \epsilon \leq 0.0273$. Then there is a subgroup $H \subseteq G$ with $dist(A, U_H) \leq 21\epsilon$.*

Proof: Pass to B and F, with

$$dist(B, B * B) \leq \epsilon.$$

A computation shows that with $\epsilon < 0.0273$, we have

$$\begin{aligned} 2\beta + 3\delta &\leq \tfrac{12}{5}\left[1 - \sqrt{1 - \tfrac{25}{6}\epsilon}\right] + O(1/N) \leq 5.16\epsilon \\ \beta + \delta &\leq 2.6\epsilon. \end{aligned}$$

From Theorem 2 we have a subset X with

$$dist(B, U_X) \leq \beta + \delta \leq 2.6\epsilon.$$

Then apply Theorem 4 with $\tau = \epsilon + 2\beta + 3\delta \leq 6.1\epsilon < 1/6$ to find the subgroup $\hat{H}$, and use the triangle inequality:

$$dist(B, U_{\hat{H}}) \leq dist(B, U_X) + dist(U_X, U_{\hat{H}}) \leq \beta + \delta + \frac{2\tau}{1 - 3\tau} < 2.6\epsilon + 3\tau \leq 21\epsilon.$$

Reverting to the original distribution,

$$dist(A, U_H) = dist(B, U_H) \leq 21\epsilon.$$

□

Once we have bounds on $dist(A, A * A)$ (or $dist(A, A \bullet A)$) and a subgroup H with small $dist(A, U_H)$, we can improve the numerical estimates of $dist(A, U_H)$.

Theorem 6. *Given a distribution A on G and a subgroup $H \subseteq G$ with*

$$dist(A, A * A) = \epsilon \leq 0.06$$

$$dist(A, U_H) = \rho \leq 0.4$$

then we can conclude

$$dist(A, U_H) \leq 5\epsilon.$$

Proof: Define $n = |H|$. Define $\rho = dist(A, U_H)$. Define $\alpha = \sum_{x \in G \setminus H} a_x$, and remark $\alpha \leq \rho$. Let $B = A \circ A$ where $\circ$ is either $*$ or $\bullet$ (or any Latin square operator respecting the subgroup H, that is, mapping $H \times H$ to H). $\sum_{x \in G \setminus H} b_x \geq 2\alpha(1 - \alpha)$, so $dist(A, B) \geq \sum_{x \in G \setminus H}(b_x - a_x) \geq 2\alpha(1 - \alpha) - \alpha = \alpha - 2\alpha^2$.

From $\alpha \leq \rho \leq 0.4$ and $\epsilon \leq 0.06$ we can compute that α is less than the larger root of $x - 2x^2 = \epsilon$, so that α is bounded by the smaller root of that equation, namely

$$\alpha \leq \frac{1 - \sqrt{1 - 8\epsilon}}{4} \leq 1.2\epsilon < 0.1.$$

Define $\psi = (1-\alpha)/n$, the average value of A on H. For $x \in H$, let $a_x = \psi + \gamma_x$, so that $\sum_H \gamma_x = 0$ and set

$$\mu = \sum_H |\gamma_x|.$$

We have

$$\mu \geq 2(\rho - \alpha).$$

This is because $2\rho = \alpha + \sum_H |\frac{\alpha}{n} - \gamma_x| \leq \alpha + \sum_H |\gamma_x| + \alpha$. We also have

$$\mu = \sum_x |\gamma_x| = \sum_H |a_x - \psi| \leq \sum_H (|a_x - 1/n| + (1/n - \psi)) \leq (2\rho - \alpha) + \alpha = 2\rho,$$

$$2\rho - 2\alpha \leq \mu \leq 2\rho.$$

For $x \in H$ define

$$Res_x = \sum_{y \circ z = x; y,z \notin H} a_y a_z.$$

Then we have

$$b_x = \sum_{y \circ z = x; y,z \in H} (\psi + \gamma_y)(\psi + \gamma_z) + Res_x$$

$$b_x = n\psi^2 + \psi(\sum_H \gamma_y + \sum_H \gamma_z) + \sum_{y,z \in H; y \circ z = x} \gamma_y \gamma_z + Res_x = n\psi^2 + \sum_{y,z \in H; y \circ z = x} \gamma_y \gamma_z + Res_x$$

because $\sum_H \gamma_y = 0$. Then

$$|b_x - \psi| \leq \frac{|\alpha - \alpha^2|}{n} + \sum_{y,z \in H; y \circ z = x} |\gamma_y \gamma_z| + Res_x$$

$$\sum_x |b_x - \psi| \leq (\alpha - \alpha^2) + \sum_{y,z \in H} |\gamma_y \gamma_z| + \alpha^2 = \alpha - \alpha^2 + \mu^2 + \alpha^2 = \alpha + \mu^2.$$

Then

$$\begin{aligned} 2dist(A,B) &= \textstyle\sum_{x \in H} |a_x - b_x| + \sum_{x \notin H} |a_x - b_x| \\ &\geq \textstyle\sum_{x \in H} |a_x - \psi| - \sum_{x \in H} |b_x - \psi| + (\alpha - 2\alpha^2) \\ &\geq \mu - (\alpha + \mu^2) + (\alpha - 2\alpha^2) \\ &= \mu - \mu^2 - 2\alpha^2 \\ \mu - \mu^2 &\leq 2\epsilon + 2\alpha^2. \end{aligned}$$

From

$$\mu \leq 2\rho \leq 2(0.4) = 0.8$$
$$2\epsilon + 2\alpha^2 < 2(0.06) + 2(0.1)^2 = 0.14,$$

we see that μ is smaller than the larger root of $x - x^2 = 2\epsilon + 2\alpha^2$, so that it is bounded by the smaller root:

$$\mu \leq \frac{1 - \sqrt{1 - 8\epsilon - 8\alpha^2}}{2} \leq \frac{1 - \sqrt{1 - 8\epsilon - 8(1.2\epsilon)^2}}{2} < 2.6\epsilon.$$

We conclude

$$dist(A, U_H) \leq 2\alpha + \mu \leq 5\epsilon.$$

□

Combining the last two results, we have:

Theorem 7. *Let A be a distribution on the finite group G. Let $dist(A, A*A) \leq \epsilon \leq 0.0273$. Then there is a subgroup $H \subseteq G$ with $dist(A, U_H) \leq 5\epsilon$.*

Example 3. Let $G = \mathbf{Z}/N$ with N a large prime integer. Let $a_n = \nu(240 - |n|)$ when $-200 < n < -1$ or $1 \leq n \leq 200$ and $a_n = 0$ otherwise, where $\nu = 1/55800$ is chosen to normalize A. Then $dist(A, A*A) \approx 0.1539$, but $dist(A, U_H) = 1 - O(1/N)$ for any subgroup H of $\mathbf{Z}/N$.

A gap remains.

4 A relationship between homomorphism testing and distributions close to their self-convolution

The two notions explored in this paper (homomorphism testing, and distributions close to their self-convolution) are related. Given a map (not necessarily a homomorphism) $f : G \to H$ between two finite groups, we can construct the product group

$$G \times H = \{(x, y) : x \in G, y \in H\}$$

and a distribution A:

$$a_{(x,y)} = \begin{cases} 1/|G| & \text{if } y = f(x) \\ 0 & \text{otherwise} \end{cases}$$

Then we have the identity

$$dist(A, A * A) = \Pr_{x,y}[f(x) \times f(y) \neq f(x \times y)].$$

If f is close to a homomorphism g, then A is close to the uniform distribution on the subgroup $\{(x, g(x)) : x \in G\}$. Conversely, if f is likely to pass the homomorphism test, or equivalently, if A is such that $dist(A, A * A)$ is small enough, then f must be close to a homomorphism: The theorem from this section implies that A must be close to a distribution that is uniform over a subgroup S of $G \times H$. One can show that S is such that for each $a \in G, |S \bigcap \{(a, b) : b \in H\}| = 1$. Then S can be used to define a map $g : G \to H$ such that g is a homomorphism (this can be seen from the underlying group structure of S) and such that g agrees with f on most inputs.

But the correspondence when $dist(A, A * A)$ is larger is not exact: the map f given in Example 1 is not at all close to any homomorphism g on all of G, but there is a subgroup $\tilde{H}$ of $G \times H$ with $dist(A, U_{\tilde{H}}) = 2/3$, namely $\tilde{H} = \{(3\ell, \ell) : 0 \leq \ell < 3^{k-1}\}$. The difference comes because g is required to be a homomorphism on all of G. We could relax the requirement, and notice that there is a large subgroup G' of G (namely $G' = \{3\ell\}$, with $|G'| = |G|/3$) and a homomorphism $g' : G' \to H$ that agrees with f on this subgroup. The subgroup $\tilde{H}$ of $G \times H$ is associated with G' and g'.

We wish to show that one can reduce the linearity question testing to the convolution on distribution question.

Theorem 8. *Let G, H be finite groups. Assume that there is a parameter β_0 and function ϕ such that the following holds:*

> *For all distributions A over group G, if $dist(A * A, A) \leq \beta \leq \beta_0$ then A is $\phi(\beta)$-close to uniform over a subgroup of G.*

Then, for any $f : G \to H$ and $\delta < \beta_0$ such that $1 - \delta = Pr[f(x) * f(y) = f(x * y)]$, and $\phi(\delta) \leq 1/2$, we have that f is $\phi(\delta)$-close to a homomorphism.

Proof: Given f, δ as above. Construct distribution A over the group $G \times H = \{(x, y) : x \in G, y \in H\}$

$$a_{(x,y)} = \begin{cases} 1/|G| & \text{if } y = f(x) \\ 0 & \text{otherwise} \end{cases}$$

Then

$$dist(A, A * A) = Pr[f(x) * f(y) \neq f(x * y)].$$

Since $\delta \leq \beta_0$, by our assumption, A is $\phi(\delta)$-close to a distribution A' which is uniform over a subgroup S of $G \times H$.

First we show that S is such that for each $a \in G$, $|S \cap \{(a, b) : b \in H\}| = 1$. Thus S can be used to define a map $g : G \to H$. Next we will show that g satisfies two properties: g is a homomorphism and g and f agree on $1 - \phi(\delta)$ fraction of the inputs in G.

Claim. S is such that for each $a \in G$, $|S \cap \{(a, b) : b \in H\}| = 1$.

Proof: S is a subgroup, so there is an integer s such that for each $a \in G$, $|S \cap \{(a, b) : b \in H\}| \in \{0, s\}$. (All those intersections have size either s or 0.) Also, the collection of a with $|S \cap \{(a, b) : b \in H\}| = s$, forms a subgroup G' of G.

We show that $s = 1$ and $G = G'$: If $G' \neq G$, then $|G'| \leq |G|/2$. In this case, $dist(A, A') \geq 1/2$ since at least half the support of A is not in A'. Thus we can assume that $G' = G$. If $s > 1$, then $dist(A, A') \geq 1/2$, since again, at least half the support of A' is missing from A. □

Because of this claim, we can define $g : G \to H$ as $g(a) = b$ for b that satisfies $(a, b) \in S$.

Claim. g is a homomorphism

Proof: For any choice of a_1, a_2, let b_1, b_2 be such that $(a_1, b_1), (a_2, b_2) \in S$, and thus, $g(a_1) = b_1, g(a_2) = b_2$. Then, since S is a subgroup, and thus closed under $+$, $(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$ is also in S, so $g(a_1 + a_2) = b_1 + b_2$.

Let e_G, e_H be the identity elements of G, H respectively. Since $(e_G, b) + (e_G, b) = (e_G, b + b) = (e_G, e_H)$, we have that $g(e_G) \to e_H$. □

Claim. g and f agree on at least $1 - \phi(\delta)$ fraction of the inputs in G.

Proof: $|x : g(x) \neq f(x)|/|G| = dist(A, A') \leq \phi(\delta)$. □

□

References

1. S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy. "Proof verification and hardness of approximation problems", *JACM*, **45(3)**: 501–555, 1998.
2. M. Bellare, D. Coppersmith, J. Hastad, M. Kiwi and M. Sudan, "Linearity Testing in Characteristic Two." FOCS 1995. *IEEE Transactions on Information Theory*, **42(6)**: 1782–1795, November 1996.
3. M. Ben Or, D. Coppersmith, M. Luby and R. Rubinfeld, "Non-Abelian Homomorphism Testing, and Distributions Close to Their Self-convolutions", in *Approximation, Randomization, and Combinatorial Optimization* (APPROX 2004 and RANDOM 2004), Springer LNCS **3122**, pp. 273–285, 2004.
4. E. Ben Sasson, M. Sudan, S. Vadhan, A. Wigderson, "Randomness-efficient Low degree tests and short PCP's via Epsilon-biased sets", in *Proceedings of the 35th STOC*, pp. 612–621, 2003.
5. M. Blum, M. Luby and R. Rubinfeld, "Self-Testing/Correcting with Applications to Numerical Problems," *J. Comp. Sys. Sci.* **47(3)**, December 1993 (special issue on STOC 1990). Preliminary abstract appears in Proc. 22th ACM Symposium on Theory of Computing, 1990.
6. D. Coppersmith, Personal communication to the authors of [5]. December 1989.
7. J. Hastad, A. Wigderson, "Simple Analysis of Graph Tests for Linearity and PCP", *Random Structures and Algorithms*, **22(2)**: 139–160, 2003.
8. A. Samorodnitsky, L. Trevisan, "A PCP characterization of NP with optimal amortized query complexity", in *Proceedings of 32nd STOC*, pp. 191–199, 2000.
9. A. Shpilka, A. Wigderson, "Derandomizing Homomorphism Testing in General Groups", in *Proceedings of 36th STOC*, pp. 427–435, 2004.