Chapter 8
Low-degree Tests

Katalin Friedl* Zsolt Hátsági† Alexander Shen‡

Abstract

We bound the Hamming distance of a multivariate function from the space of polynomials of max-degree $\leq k$ in terms of its distances from functions that are polynomials of degree $\leq k$ in one of their variables. The latter quantities can be estimated easily by statistical tests. The result conceptualizes an algorithm analysis of M. Szegedy and incorporates improvements based on a result of S. Arora and S. Safra. The bound results in a query-efficient BFL/FGLSS-type max-degree test, an ingredient of transparent proofs. We also describe another variant of the BFL max-degree test which works over smaller domains. The analysis of this version follows the ideas of [12] and uses a combinatorial isoperimetric inequality.

1 Introduction

Testing that a multivariate function, given by an array of its values, is approximately a low-degree polynomial, plays a central role in a series of papers on interactive and transparent proofs ([5], [6], [12], [13], [3], [2], [9], [17]). The *objective* is to accept the array if it represents a low-degree polynomial, and reject it with large probability if it is not close (in Hamming distance) to some low-degree polynomial. The decision should be based on a small number of randomized queries to the array.

Low-degree tests come in two brands: *max-degree* tests and *total-degree* tests. While max-degree tests have been considered over arbitrary finite subsets of a field, the known total-degree tests require the domain to be the entire (necessarily finite) field. Considering arbitrary subsets of a field allows the use of the field of rationals, as was required in [5] and [6].

For low-degree tests, two minimization goals have been considered: minimization of the size of the domain over which the test works, and minimization of the number of queries required by the test. The former is directly related to the length of the transparent proof, the latter to the number of spot-checks needed to verify the transparent proof.

The first low-degree test appears in the paper of Babai, Fortnow, and Lund [5] which proves that multiple prover interactive protocols have the power of nondeterministic exponential time. All subsequent max-degree tests are variants of that test. There are, however, differences in the analysis. The BFL-test [5] was analysed via a *combinatorial isoperimetric inequality*. A stronger isoperimetric inequality is used by Feige, Goldwasser, Lovász, Safra, and Szegedy [12]. A variant of that analysis was used to achieve optimal domain size in [7], as required by the transparent proofs of Babai, Fortnow, Levin, and Szegedy [6].

More recent papers of Arora and Safra [3], Arora, Lund, Motwani, Sudan, and Szegedy [2] have used total-degree tests which are more efficient in terms of the number of queries and ultimately also matched the domain size of the best max-degree tests (cf. Sudan [21]).

Nevertheless, we believe that further clarification of max-degree tests is a worthwhile task. In particular, the fact that these tests do not require the domain to be the entire field may have future applications. A re-examination of certain max-degree tests is the subject of the present paper.

An early version of [12] included an unpublished analysis of the max-degree test by Szegedy (which is different from the analysis that eventually appeared in [12]). That analysis has been the starting point of our work.

By separating out the mathematical contents of that analysis, we are able to give a more transparent analysis, as well as incorporate further improvements.

The main results are two inequalities providing explicit upper bounds on the Hamming distance of an n-variate function from the space of polynomials of max-degree $\leq k$ in terms of its distances from functions that are polynomials of degree $\leq k$ in one of their

*Department of Computer Science, University of Chicago, and Hungarian Academy of Sciences. Research supported in part by NSF Grant CCR-9014562 and OTKA Grant 2581 (Hungary). e-mail: friedl@cs.uchicago.edu

†Department of Computer Science, University of Chicago, and Agricultural Biotechnology Center, Hungary. e-mail: hatsagi@cs.uchicago.edu

‡Institute of Problems of Information Transmission, Lab 13, Ermolovoi 19, Moscow 101447, Russia. This author is grateful to the A.M.S. and the Moscow Mathematical Institute for support. e-mail: shen@sch57.msk.su

variables. The latter quantities can be estimated easily by statistical tests. The first of the two bounds holds even for small domains. Using a recent result of Arora and Safra [3] instead of a trivial estimate gives an order of magnitude improvement in the main term and an improved error term, but this argument seems to require larger domain size ($\Omega(nk^3)$).

The analysis of the FGLSS max-degree test [12] will now be an immediate consequence. By the first result, $O(nk^2)$ queries to the values of f suffice to give error probability less than $1/2$, assuming domain size $\Omega(n^2k^2)$. The second result implies that $O(nk)$ queries suffice, assuming domain size $\Omega(n^2k + nk^3)$.

For completeness, we also include an unpublished variant of the BFL max-degree test given in [7] which achieves the smallest domain sizes (domain size $\geq 4nk+2$) and requires $O(n^2k^2)$ queries when the domain size is $\Theta(nk)$.

For comparison, we mention some results for total-degree tests (total degree $\leq d$). There are methods requiring $O(d^2)$ queries over domains of size $\Omega(nd)$ or $O(d)$ queries over domains of size $\Omega(nd^2)$ ([19], [21], [14], [2]).

2 Main results

DEFINITION 2.1. *Let F be a field and $I \subseteq F$ a finite subset of F.*
A polynomial of n variables over F is a max-degree-k *polynomial if its degree with respect to each variable is not greater than k.*
A function $f : I^n \to F$ is called a max-degree-k polynomial, if it has a max-degree-k extension to F^n.

Let $P(n,k)$ be the set of max-degree-k polynomials of n variables over I and $P_i(n,k)$ be those $I^n \to F$ functions that are polynomials of degree $\leq k$ with respect to the ith variable (i.e. f has an extension to F^n with this property). Then clearly $P(n,k) = \cap_{i=1}^n P_i(n,k)$.

DEFINITION 2.2. *The* distance *(normalized Hamming distance) of two functions, $f, g : I^n \to F$ is*

$$d(f,g) = Prob_{x \in I^n}(f(x) \neq g(x)).$$

(The probability is with respect to the uniform distribution over I^n.)

The goal is to estimate how far a function is from being a max-degree-k polynomial assuming it is close to a degree-k univariate polynomial in the i-th variable (for any fixed value of the other variables) for each i.

THEOREM 2.1. *For any function $f : I^n \to F$ and for any k*

$$(2.1)\quad d(f,P(n,k)) \leq 6(k+1)\sum_{i=1}^{n} d(f,P_i(n,k)) + 2\frac{nk}{\sqrt{|I|}}.$$

This estimate is vacuously true for domains of size $|I| \leq 9n^2k^2$, (the error term is at least 1), but it gives a meaningful bound for $|I| = \Omega(n^2k^2)$ and this case will be used in the analysis of the algorithm.

For larger domains ($\Omega(nk^3)$), both the main term and the error term can be reduced.

THEOREM 2.2. *For any function $f : I^n \to F$ and for any k with $|I| > 18nk^3$,*

$$(2.2)\quad d(f,P(n,k)) \leq 12\sum_{i=1}^{n} d(f,P_i(n,k)) + 3\frac{n\sqrt{k}}{\sqrt{|I|}}.$$

3 Max-degree test

Both theorems lead to an analysis of the parameters of the FGLSS max-degree test.

DEFINITION 3.1. *We say that a max-degree-k test is* ε-reliable *if*

(i) the test always accepts if f is a max-degree-k polynomial,

(ii) if $d(f,P(n,k)) > \varepsilon$ then the probability of acceptance is $\leq 1/2$.

By selecting an element of a finite set S "at random" we mean a random variable, uniformly distributed over S.

The algorithm consists of repeating the following Line-test m times.

Line-test

- Fix elements $a_1, \ldots a_{k+1} \in I$.
- Choose i at random from $\{1, \ldots, n\}$.
- Pick independent random elements $u_1, \ldots, u_n \in I$.
- Check if the restriction of f to the $k+2$ points

$$(u_1, \ldots, u_{i-1}, u_i, u_{i+1}, \ldots, u_n),$$
$$(u_1, \ldots, u_{i-1}, a_j, u_{i+1}, \ldots, u_n), \qquad 1 \leq j \leq k+1;$$

determines a degree-k polynomial (as a univariate function of variable x_i). If not, reject.

For the analysis assume that we have an estimate for the distance of the function f from the set of max-degree-k polynomials in the form

$$(3.3) \qquad d(f, P(n,k)) \le a \sum_{i=1}^{n} d(f, P_i(n,k)) + \Delta$$

for some a and Δ.

Let $d(f, P(n,k)) \ge \varepsilon$. The probability that in a round when i was chosen f is rejected is at least $d(f, P_i(n,k))$. Therefore the probability that f does not pass one round with a random i is at least $\sum_i d(f, P(n,k))/n \ge (\varepsilon - \Delta)/(an)$.

Number of steps and number of random bits used. If the function is given by an array of its values (with random access) then the number of queries is $m(k+2) = O(akn/\varepsilon)$ and $O(akn/\varepsilon)$ arithmetic operations are performed in F.

The number of random bits used is a key resource. Successive improvements of the transparent proof technique and its applications to characterizing NP and proving intractability of approximate discrete optima ([5], [12], [3], [2], [9]) critically depended on reducing this number.

One round uses $\log n + n \log |I| = O(n \log |I|)$ random bits. Repeating this m times would require $O(mn \log |I|)$ random bits. However, as pointed out in [12], it suffices to select the tuples $(i, x_1, \ldots, x_n) \in \{1, \ldots, n\} \times I^n$ pairwise independently. Up to $n|I|^n$ pairwise independent tuples can be generated from only two fully independent ones ([18], [16], cf. [1], [11]) thus the number of random bits needed is still only $O(n \log |I|)$.

COROLLARY 3.1. *Assume $|I| > 16n^2k^2/\varepsilon^2$. Set $m = O(nk/\varepsilon)$. Then an m-fold pairwise independent repetition of the Line-test constitutes an ε-reliable max-degree-k test which makes $O(nk^2/\varepsilon)$ queries and uses $O(n \log n)$ random bits.*

Proof. By Theorem 2.1, in inequality (3.3) the parameters are $a = 3(k+1)$ and $\Delta = 2nk/\sqrt{|I|}$. The condition on $|I|$ implies that $\Delta < \varepsilon/2$. □

COROLLARY 3.2. *Assume $|I| > 18nk^3$ and $|I| > 32n^2k/\varepsilon^2$. Set $m = O(n/\varepsilon)$. Then an m-fold pairwise independent repetition of the Line-test constitutes an ε-reliable max-degree-k test which makes $O(nk/\varepsilon)$ queries and uses $O(n \log n)$ random bits.*

Proof. By Theorem 2.2 in inequality (3.3) the parameters are $a = 12$ and $\Delta = 3n\sqrt{k}/\sqrt{|I|}$. The condition on $|I|$ implies that $\Delta < \varepsilon/2$. □

4 Distance estimates

A result of J. T. Schwartz [20] and R.E. Zippel [23] gives a tool to estimate the distance of two polynomials:

LEMMA 4.1. (SCHWARTZ, ZIPPEL)
Let $g(x_1, \ldots, x_n)$ be a nonzero polynomial over a field F. Let d denote its total degree. For any finite subset $I \subseteq F$, if $a_i \in I$ are chosen independently uniformly at random then $Prob(g(a_1, \ldots, a_n) = 0) \le d/|I|$.

It follows that the distance of any two polynomials, $g, h \in P(n,k)$ is large: if $g \ne h$ then $d(g,h) \ge 1 - kn/|I|$. Consequently for any function f, if g is a polynomial from $P(n,k)$ such that $d(f,g) < (1 - kn/|I|)/2$ then g is the (unique) closest polynomial to f and $d(f, P(n,k)) = d(f,g)$. We call the distance $d(f, P(n,k))$ the *degree-k error* of f.

The next question we consider is the connection between the degree-k error of an n-variate function and its restrictions. Clearly

$$(4.4) \qquad d(f, P(n,k)) \ge \frac{1}{|I|} \sum_{c \in I} d(f|_{x_1 = c}, P(n-1,k)).$$

Equality means that, when $g \in P(n,k)$ denotes a closest max-degree-k polynomial to f, the restriction of g to any of the hyperplanes $x_1 = c$, $c \in I$ gives a closest polynomial on that hyperplane to the restriction of f.

When f is not just any function but it is a polynomial of degree $\le k$ in its first variable then we can say more. As we shall see, in this case strict inequality implies that for most hyperplanes $x_1 = c$ the distance there is large (consequently $d(f, P(n,k))$ is also large); on the other hand if equality holds and $d(f, P(n,k))$ is not too large then the distance on most hyperplanes is close to $d(f, P(n,k))$.

For the rest of this section let $f \in P_1(n,k)$. Let $g \in P(n,k)$ be a closest polynomial to f, $d(f, P(n,k)) = d(f,g)$. We call a point $x \in I^n$ *good* with respect to g if $f(x) = g(x)$ and *bad* otherwise. By *lines in the i^{th} direction* we shall mean lines parallel to the x_i axis, i. e. sets of the following type

$$\{(u_1, \ldots, u_{i-1}, z, u_i, \ldots, u_n) \mid z \in I\}$$

for fixed $u_1, \ldots, u_n \in I$. (There are $|I|^{n-1}$ lines in each of the n directions.) A line is called *bad* if somewhere on that line f and g disagree. By the "proportion of bad lines in the first direction" we mean the proportion of bad lines among the lines in the first direction. Let α denote this proportion. Note that what counts as a bad point or bad line depends on g.

LEMMA 4.2. *If the proportion of bad lines in the first direction with respect to some $g \in P(n,k)$ is*

$\alpha < 1/3$ then g is the unique element of $P(n,k)$ closest to f and equality holds in (4.4).

Proof. Clearly

$$d(f,g) \le \alpha$$

and

$$(\forall c \in I) \qquad d(f|_{x_1=c}, g|_{x_1=c}) \le \alpha.$$

The assumption that $\alpha \le 1/3$ implies that g is the closest max-degree-k polynomial to f and also, $g|_{x_1=c}$ is the closest max-degree-k polynomial to $f|_{x_1=c}$ for any $c \in I$, so

$$d(f,P(n,k)) = d(f,g) = \frac{1}{|I|}\sum_{c\in I} d(f|_{x_1=c}, g|_{x_1=c}) = \frac{1}{|I|}\sum_{c\in I} d(f|_{x_1=c}, P(n-1,k)).$$

□

The following two lemmas will imply that when there is strict inequality in (4.4) then the degree-k errors on most of the hyperplanes $x_1 = c$ are large.

LEMMA 4.3. *Let $f \in P_1(n,k)$ and $|I| > 3nk$. Assume that for some $\tau \le 1/(3k+3)$,*

$$|\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) \le \tau\}| \ge k+1.$$

Then the proportion of bad lines in the first direction is $\alpha \le \tau(k+1)$.

Proof. Let $d(f_{x_1=c_i}, P(n-1,k)) \le \tau$ for some $c_1, \ldots, c_{k+1} \in I$ and let $g_i \in P(n-1,k)$ denote a closest max-degree-k polynomial to $f|_{x_1=c_i}$ (on the hyperplane $x_1 = c_i$), i.e. $d(f_{x_1=c_i}, P(n-1,k)) = d(f_{x_1=c_i}, g_i)$. Let g be the degree-k extension of the g_i to the entire I^n. Then $g \in P(n,k)$. We shall show that g is the closest element of $P(n,k)$ to f.

Both f and g are polynomials of degree at most k on any line in the first direction. On a bad line in the first direction they differ everywhere except at most k places. Clearly α is an upper bound for $d(f,g)$ and also for $d(f|_{x_1=c}, g|_{x_1=c})$ for any $c \in I$, since the points where f and g are different are covered by the bad lines.

If a line intersects each of the $k+1$ hyperplanes $x_1 = c_i$ in a good point then f and g agree on the whole line. On any of these hyperplanes the proportion of the points which are not good is at most τ, therefore the proportion of bad lines is $\le \tau(k+1) \le 1/3$. Now since $d(f,g) \le \alpha \le 1/3 < (1 - k/|I|)/2$, Lemma 4.1 implies that g is the unique closest element of $P(n,k)$, and so $\alpha \le \tau(k+1)$. □

LEMMA 4.4. (ARORA, SAFRA) *Let $f \in P_1(n,k)$ and $|I| > 18nk^3$. Assume that for some $\tau \le 1/6$*

$$|\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) \le \tau\}| \ge 2k.$$

Then the proportion of bad lines in the first direction is $\alpha \le 2\tau$.

COROLLARY 4.1. *Let $f \in P_1(n,k)$ and $|I| > 3nk$. If*

$$d(f,P(n,k)) > \frac{1}{|I|}\sum_c d(f|_{x_1=c}, P(n-1,k))$$

then

$$d(f,P(n,k)) > \frac{2}{9(k+1)}.$$

If $|I| > 18nk^3$ then

$$d(f,P(n,k)) > \frac{1}{7}.$$

Proof. When strict inequality holds in (4.4) then by Lemmas 4.3 and 4.2,

$$|\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) \le 1/(3k+3)\}| < k+1.$$

Therefore

$$d(f,P(n,k)) > \left(1 - \frac{k}{|I|}\right)\frac{1}{3k+3} \ge \frac{2}{9k+9}.$$

The second claim follows similarly from Lemmas 4.4 and 4.2. □

Even if equality holds in (4.4) we can say something about the distances $d(f|_{x_1=c}, P(n-1,k))$. Not just that their average is exactly $d(f,P(n,k))$ but for most $c \in I$, $d(f|_{x_1=c}, P(n-1,k)$ cannot be much smaller then $d(f,P(n,k))$.

LEMMA 4.5. *Let $f \in P_1(n,k)$ and assume that equality holds in (4.4). Then for $0 < \mu < 1$*

$$(4.5)\quad |\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) < d(f,P(n,k)) - \mu)\}| \le k\alpha/(\mu|I|).$$

Proof. The proportion of bad lines in the first direction, α is an upper bound for $d(f,P(n,k))$ and $d(f|_{x_1=c}, P(n-1,k))$ for any $c \in I$. These distances differ from α because there can be good points on the bad lines. On a line in the first direction both f and g are polynomials of degree at most k, so if they are different they agree in at most k points. Therefore the total number of good points on bad lines is $\le k\alpha|I|^{n-1}$. □

5 Proofs of the theorems

The proofs of the two theorems are the same, the difference is whether Lemma 4.3 or Lemma 4.4 is applied during the proof. To be able to prove the two theorems together here is a parameterized form of the lemmas:

Let $f \in P_1(n\ k)$. Then for some t and τ either

$$(5.6)\quad |\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) \le \tau\}| \le t.$$

or with $\mu = \sqrt{k\tau}/\sqrt{3|I|}$

$$(5.7)\quad |\{c \in I \mid d(f|_{x_1=c}, P(n-1,k)) \ge d(f, P(n,k)) - \mu) \le \mu/\tau$$

This will imply that for any function f

$$(5.8)\quad d(f, P(n,k)) \le \frac{2}{\tau}\sum_{i=1}^{n} d(f, P_i(n,k)) + 2n\frac{\mu}{\tau}.$$

Obse rve that the values $\tau = 1/(3k+3)$, $t = k$, $\mu \not\le k/\sqrt{9(k+1)|I|}$ guaranteed by Lemma 4.3, results in Theorem 2.1. From values $\tau = 1/6$, $t = 2k-1$, $\mu \sqrt{k}/\sqrt{18|I|}$ (Lemma 4.4) we obtain Theorem 2.2.

So what is left is to prove inequality (5.8) from (5.7) and (5.6).

Let $f^i \in P_i(n,k)$ be closest to f, i.e.

$$d(f, f^i) = d(f, P_i(n,k)).$$

The proof goes by induction showing the following

CLAIM. For $1 \le i \le n$ and for any $(c_1, \ldots, c_i) \in I^i$, except a fraction of $(i \max\{\mu/\tau, t/|I|\})$ the following holds:

$$(5.9)\quad \tau\, d(f, P(n,k)) \le i\mu + d(f, f^1) + \sum_{j=1}^{i-1} d(f^j|_{x_1=c_1,\ \ldots,x_j=c_j}, f^{j+1}|_{x_1=c_1,\ \ldots,x_j=c_j}) + d(f^i|_{x_1=c_1,\ \ldots,x_i=c_i}, P(n-i,k)).$$

Proof of Claim We proceed by induction on i.

For the function f^1 either (5.6) or (5.7) holds. In the first case

$$\tau d(f, P(n,k)) \le \tau < d(f^1|_{x_1=c_1}, P(n-1,k))$$

except for a fraction $\le t/|I|$ of the $c_1 \in I$. In the second case by the triangle inequality

$$d(f, P(n,k)) \le d(f, f^1) + d(f^1, P(n,k))$$

and

$$d(f^1, P(n,k)) < d(f^1|_{x_1=c_1}, P(n-1,k)) + \mu$$

is true with μ/τ exceptions.

Similarly for any $1 \le i < n-1$ if (5.6) holds for the function $f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i}$ then

$$\tau d(f, P(n,k)) \le \tau < d(f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i,x_{i+1}=c_{i+1}}, P(n-i-1,k))$$

except for a fraction $\le t/|I|$ of the $c_{i+1} \in I$. If (5.7) is true then by the triangle inequality

$$d(f^i|_{x_1=c_1,\ \ldots,x_i=c_i}, P(n-i,k)) \le d(f^i|_{x_1=c_1,\ \ldots,x_i=c_i}, f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i}) + d(f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i}, P(n-i,k))$$

and

$$d(f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i}, P(n-i,k)) \le \mu + d(f^{i+1}|_{x_1=c_1,\ \ldots,x_i=c_i,x_{i+1}=c_{i+1}}, P(n-i-1,k))$$

holds except for a fraction μ/τ of the $c_{i+1} \in I$. This proves the Claim. □

For both of our settings, $t/|I| < \mu/\tau$. The exceptions in the Claim are a fraction of $(n \max\{\mu/\tau, t/|I|\}) \le n\mu/\tau$ of the $(c_1, \ldots, c_n) \in I^n$. For a fixed $(c_1, \ldots, c_n) \in I^n$ consider either inequality (5.9) or for the exceptions the inequality where the right hand side of (5.9) is increased by τ (these modified inequalities always hold, since the left hand side is at most τ). Now take the average of these inequalities over all possible choices of $(c_1, \ldots, c_n) \in I^n$. This gives

$$\tau\, d(f, P(n,k)) \le 2n\mu + d(f, f^1) + \sum_{j=1}^{n-1} d(f^j, f^{j+1}).$$

Apply the triangle inequality

$$d(f^j, f^{j+1}) \le d(f, f^j) + d(f, f^{j+1})$$

in the sum for $1 \le j \le n-1$. Then (5.8) follows. □

As it was shown after (5.8) this finishes the proofs of Theorem 2.1 and 2.2. □

6 Small domains

The following variant of the BFL max-degree test is described in [7].

Full-line-test

- Choose i at random from $\{1, \ldots, n\}$.
- Pick a random line ℓ in direction i.
- Check if the restriction of f to ℓ determines a degree-k polynomial (as a univariate function of variable x_i). If not, reject.

THEOREM 6.1. *Assume* $|I| \geq 4nk + 2$. *Set* $m = O(n(k + 1/\varepsilon))$. *Then an* m*-fold pairwise independent repetition of the Full-line-test constitutes an* ε*-reliable max-degree-*k *test which makes* $O(|I|n(k+1/\varepsilon))$ *queries and uses* $O(n \log n)$ *random bits.*

Note that for the minimum domain size $|I| = 4nk + 2$, the number of queries is $O(n^2k^2 + n^2k/\varepsilon)$. This seems to be the only known max-degree test allowing domain size *linear* in k, as required for transparent proofs of nearly linear size [6].

Theorem 6.1 is a corollary to the following result. Let W denote the set of lines ℓ (in any direction) such that f restricted to ℓ is not a polynomial of degree $\leq k$. Call the elements of W *wrong lines*. Let $\beta = |W|/(n|I|^{n-1})$ denote the proportion of wrong lines.

THEOREM 6.2. ([7]) *For any function* $f : I^n \to F$ *and for any* k *with* $|I| \geq 4nk + 2$ *either*
(a) the proportion of wrong lines is

$$\beta > \frac{1}{4n(k+1)}$$

(b) or

$$d(f, P(n,k)) \leq n\beta.$$

The proof follows the ideas of [12] and is based on a combinatorial isoperimetric inequality, appearing in [8]. For completeness, we include the proof.

Let $G(V, E)$ be a graph and $B \subseteq V$.

Let $\delta(B)$ denote the set of edges $\{v, w\} \in E$ such that $v \in B$, $w \notin B$.

A graph is *edge-transitive* if all edges are equivalent under automorphisms (self-isomorphisms).

PROPOSITION 6.1. ([8]) *Let* G *be a connected edge-transitive graph and* $B \subseteq V$ *a nonempty subset such that* $|B| \leq |V|/2$. *Then*

$$|\delta(B)|/|B| \geq r/(2diam(G)), \tag{6.10}$$

where $diam(G)$ *denotes the diameter of* G *and* r *is the minimum degree.*

(We remark that the result holds even with r denoting the harmonic mean of the minimum and the maximum degrees of G.)

In our application, the graph G to be considered is the Hamming-graph on I^n: its vertex set is I^n and two points are adjacent if their Hamming distance is 1 (differ in a single coordinate). The degree of each vertex is $r = n(|I| - 1)$; the diameter is $\mathrm{diam}(G) = n$.

LEMMA 6.1. *Let* $|I| \geq 4nk + 2$. *If* $d(f, P(n,k)) \leq 1/2$ *then* $d(f, P(n,k)) \leq n\beta$.

Proof. By assumption, $d(f, g) \leq 1/2$ for some $g \in P(n,k)$. Let $B = \{x \in I^n; f(x) \neq g(x)\}$; call the elements of B *bad points.*

Let us call a line ℓ *deceptive* if the restriction $f|_\ell$ is a polynomial of degree $\leq k$ in the corresponding variable, but $f|_\ell \neq g|_\ell$. Let D denote the set of deceptive lines. First we observe that any deceptive line contains at least $|I| - k$ bad points (since two univariate polynomials of degree $\leq k$ cannot agree at more than k places). It follows that

$$n|B| \geq |W| + (|I| - k)|D|. \tag{6.11}$$

Indeed, in any direction i ($1 \leq i \leq n$), each wrong line contains at least one bad point, and any deceptive line contains at least $|I| - k$ bad points. Adding these counts up for all i, we counted each bad point at most n times.

Let us now count the edges of G leaving B. Each such edge determines a line which is either wrong or deceptive. A line cannot contribute more than $|I|^2/4$ to $\delta(B)$; and a deceptive line contributes at most $k(|I| - k)$ (since on such a line, all but at most k points are bad). To sum up, we have

$$|\delta(B)| \leq (|I|^2/4)|W| + k(|I| - k)|D|.$$

Combining this with the isoperimetric inequality 6.10 with $r = n(|I| - 1)$ and $\mathrm{diam}(G) = n$, we obtain

$$(|I| - 1)|B|/2 \leq (|I|^2/4)|W| + k(|I| - k)|D|.$$

Expressing $(|I| - k)|D|$ from inequality (6.11), we infer

$$(|I| - 1)|B|/2 \leq (|I|^2/4)|W| + k(n|B| - |W|).$$

Rearranging and taking $|I| \geq 4nk + 2$ into account, we obtain

$$|I| \cdot |B| \leq (2(|I|-1) - 4nk)|B| \leq (|I|^2 - 4k)|W| \leq |I|^2|W|.$$

Dividing the two extreme sides by $|I|^{n+1}$, we obtain

$$d(f, P(n,k)) = \frac{|B|}{|I|^n} \leq n\beta.$$

□

The following Lemma, combined with the previous one, will complete the proof of the Theorem.

LEMMA 6.2. *Let $|I| \geq 4nk+2$. If $d(f, P(n,k)) \geq 1/(3(k+1))$ then $\beta \geq (1-k/|I|)^n/(3n(k+1))$.*

Proof. We proceed by induction on n; the case $n=1$ is clear. Let now $n \geq 2$.

Case 1. Assume that for all but $\leq k$ values of $c \in I$,

$$d(f|_{x_n=c}, P(n-1,k)) \geq 1/(3(k+1)).$$

By the inductive hypothesis, the proportion of wrong lines in each of these hyperplanes is $\geq (1-k/|I|)^{n-1}/(3(n-1)(k+1))$. Adding these up we obtain

$$\beta \geq \frac{|I|-k}{|I|} \cdot \frac{n-1}{n} \cdot \frac{(1-k/|I|)^{n-1}}{3(n-1)(k+1)} = \frac{(1-k/|I|)^n}{3n(k+1)},$$

as desired.

Case 2. There exist $k+1$ distinct values $c_1, \ldots, c_{k+1} \in I$ such that

$$d(f|_{x_n=c_j}, P(n-1,k)) < 1/(3(k+1)).$$

Let $g_j(x_1, \ldots, x_{n-1}) \in P(n-1,k)$ realizing the Hamming distance $d(f|x_n = c_j, P(n-1,k))$. These $(n-1)$-variate polynomials determine a unique polynomial $g \in P(n,k)$ such that $g_j = g|_{x_n=c_j}$.

Let $B_j \subset I^{n-1}$ denote the set of bad points for $f|_{x_n=c_j}$. Let W_n denote the set of wrong lines (for f) in the n^{th}-direction.

If $|W_n| \geq |I|^{n-1}/(3(k+1))$ then

$$\beta \geq |W_n|/(n|I|^{n-1}) \geq 1/(3n(k+1))$$

and we are done. On the other hand, if If $|W_n| < |I|^{n-1}/(3(k+1))$ then we claim that

$$d(f,g) < 1/2.$$

Indeed, let again $B = \{x \in I^n; f(x) \neq g(x)\}$ (bad points). Now if x is bad then either $x \in \ell$ for some $\ell \in W$, or $x = (y,c)$ for some $y \in \bigcup_{j=1}^{k+1} B_j$. Now,

$$|\bigcup_{\ell \in W_n} \ell| = |I||W_n| < |I|^n/(3(k+1)).$$

Moreover, $|B_j| \leq |I|^{n-1}/(3(k+1))$ for each j, so each B_j accounts for at most $|I|^n/(3(k+1))$ bad points. Adding this all up, we obtain

$$|B|/|I|^n \leq 1/3 + 1/(3(k+1)) \leq 1/2.$$

By Lemma 6.1, we infer that

$$\beta \geq \frac{1}{n} d(f, P(n,k)) \geq \frac{1}{n} \frac{1}{3(k+1)}.$$

□

Proof of Theorem 6.2. If $d(f, P(n,k)) \leq 1/2$, conclusion (b) follows by Lemma 6.1. If $d(f, P(n,k)) > 1/2$ then certainly $d(f, P(n,k)) > 1/(3(k+1))$ and we conclude by Lemma 6.2 that

$$\beta \geq (1-k/|I|)^n/(3n(k+1)) > 1/(4n(k+1))$$

(using the condition that $|I| > 4nk$). □

7 Open problems

The main question is how to reduce simultaneously the size of the domain and the number of queries in the max-degree test.

Observe, that for example a better estimate for the distance of f from the max-degree-k polynomials in the form (3.3) can help. An improvement in the error term Δ would allow one to use smaller domain in the algorithm. A smaller constant a would reduce the number of rounds, i.e. the number of queries. We do not even know if a formula of type (3.3) may hold with $\Delta = 0$.

Acknowledgment

We are grateful to Mario Szegedy for allowing us to elaborate on his unpublished work. We thank Laci Babai for comments and encouragement.

References

[1] N. Alon, L. Babai, and A. Itai, *A fast and simple randomized parallel algorithm for the maximal independent set problem*, J. of Algorithms, 7 (1986), pp. 567–583.

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, *Proof verification and hardness of approximation problems*, In: Proc. 33rd FOCS, IEEE 1992, pp. 14–23.

[3] S. Arora and S. Safra, *Probabilistic checking of proofs; a new characterization of NP*, In: Proc. 33rd FOCS, IEEE 1992, pp. 2–13.

[4] L. Babai, *Transparent proofs and limits to approximation*, In: Proc. First European Congress of Mathematics, Vol.1., Birkhäuser, to appear.

[5] L. Babai, L. Fortnow, and C. Lund, *Non-deterministic exponential time has two-prover interactive protocols*, Computational Complexity 1 (1991), pp. 41–66.

[6] L. Babai, L. Fortnow, L. Levin, and M. Szegedy, *Checking computations in polylogarithmic time*, In: Proc. 23rd STOC, ACM 1991, pp. 21–31.

[7] L. Babai and K. Friedl, *On slightly superlinear transparent proofs*, Tech. Report CS 93-13, Department of Computer Science, University of Chicago, 1993.

[8] L. Babai and M. Szegedy, *Local expansion of symmetrical graphs*, Combinatorics, Probability and Computing 1 (1992), pp. 1–11.

[9] M. Bellare, S. Goldwasser, C. Lund, and A. Russel, *Efficient Probabilistic checkable proofs and applications*

to approximation, In: Proc. 25nd STOC, ACM 1993, pp. 294–304.

[10] E. Berlekamp and L. Welch, *Error correction of algebraic block codes*, US Patent Number 4,633,470.

[11] B. Chor, O. Goldreich, J. Hastad, J. Friedman, S. Rudich, and S. Smolensky, *t-resilient functions* In: Proc. 26th FOCS, IEEE 1985, pp. 396–407.

[12] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy, *Approximating clique is almost NP-complete*, In: Proc. 32nd FOCS, IEEE 1991, pp. 2–12.

[13] U. Feige and L. Lovász, *Two-prover one-round proof systems: their power and their problems*, In: Proc. 24nd STOC, ACM 1992. pp. 733–744.

[14] P. Gemmell, R. Lipton, R. Rubinfeld, M. Sudan, and A. Wigderson, *Self-testing/correcting for polynomials and for approximate functions*, In: Proc. 23rd STOC, ACM 1991, pp. 32–43.

[15] P. Gemmell and M. Sudan, *Highly resilient correctors for polynomials*, Inf. Proc. Letters 43 (1992), pp. 169–174.

[16] A. Joffe, *On a set of almost deterministic k-independent random variables* Annals of Prob. 13 (1992), pp. 502–524.

[17] J. Kilian, *A note on efficient zero-knowledge proofs and arguments*, In: Proc. 24th STOC, ACM 1992, pp. 723–732.

[18] H.O. Lancaster, *Pairwise statistical independence*, Ann. Math. Stat., 36 (1965), pp. 1313–1317.

[19] R. Rubinfeld and M. Sudan, *Testing polynomial functions efficiently and over rational domains*, In: Proc. 3rd SODA, ACM–SIAM 1992, pp. 23–32.

[20] J.T. Schwartz, *Fast probabilistic algorithms for verification of polynomial identities*, J. of ACM 27 (1980), pp. 701–717.

[21] M. Sudan, *Efficient Checking of polynomials and Proofs and the Hardness of Approximation Problems*, PhD thesis, University of California at Berkeley, 1992.

[22] M. Szegedy, *Multilinearity test*, unpublished, 1991.

[23] R.E. Zippel, *Probabilistic algorithms for sparse polynomials*, In: Proc. EUROSAM'79, Lecture Notes in Computer Science 72, 1979, pp. 216–226.