1 The degree-h test

Lecture 4: April 9, 2003

We use $|\mathcal{F}|$ to denote the number of distinct elements in the finite field $\mathcal{F}$. We assume here that $|\mathcal{F}|$ is a prime and the members of $\mathcal{F}$ are the numbers from 0 to $(|\mathcal{F}|-1)$. The *distance* between two functions $f_1, f_2 : \mathcal{F}^m \to \mathcal{F}$,

$$\Delta(f_1, f_2) = \frac{|\{\vec{y} | f_1(\vec{y}) \neq f_2(\vec{y})\}|}{|\mathcal{F}|^m}$$

is the fraction of $\mathcal{F}^m$ on which f_1 and f_2 disagree. For $h < |\mathcal{F}|$, let DH be the set of degree-h multinomials defined on $\mathcal{F}^m$. We define the *minimal distance* between a function $f : \mathcal{F}^m \to \mathcal{F}$ and a degree-h multinomial

$$\Delta_{DH}(f) = \min_{f' \in DH} \Delta(f, f').$$

We call a set of $|\mathcal{F}|$ points $\left\{\vec{y}_1, \ldots, \vec{y}_{|\mathcal{F}|}\right\} \subseteq \mathcal{F}^m$ an *aligned line* in direction i if they differ only in the i^{th} coordinate.

Let f be a function from $\mathcal{F}^m$ to $\mathcal{F}$. It induces the following m functions, $f_1, \ldots, f_m$, where f_i for $1 \leq i \leq m$ is defined as follows. For every aligned line in direction i, the values of f_i on the first $h+1$ points on the line is equal to the respective values of f on these points. (Namely, for every point $\vec{y} = (y_1, \ldots, y_m) \in \mathcal{F}^m$ for which $y_i \leq h$, $f_1(\vec{y}) = f(\vec{y})$.) There is a unique degree h univariate polynomial $p(x)$ that agrees with these values on these $h+1$ points. (Moreover, recall that such a polynomial can be found in time polynomial in the input size.) The values of f_i for all other points on the aligned line are determined by evaluating the polynomial $p(x)$ on the respective point y_i.

The basic degree-h test. Pick a random point $\vec{y} \in \mathcal{F}^m$ and a random coordinate $i \in \{1, \ldots, m\}$. Accept if and only if $f(\vec{y}) = f_i(\vec{y})$.

The number of random bits needed in order to perform the basic degree-h test is $m \log |\mathcal{F}|$. The number of field elements that one needs to read from f is $h+2$.

Lemma 1 *f is a degree-h multinomial iff $f = f_i$ for every $1 \leq i \leq m$. (Other equivalent conditions are iff the basic test accepts with probability 1, and iff f is a degree h univariate polynomial over every aligned line.)*

Proof: If f is a degree-h multinomial then restricting it on an aligned line (meaning, giving values to all variables except y_i) gives a univariate polynomial in y_i of degree h, which must be identical to the respective polynomial $p(x)$ that defines f_i, because both polynomials agree on their first $h+1$ values.

Conversely, assume that $f = f_i$ for all $1 \leq i \leq m$. Let g be the unique degree-h multinomial that agrees with f on all points $\vec{y} \in [0, h]^m$. (We have seen in lecture 3 that such a multinomial exists.) We need to show that g agrees with f on all $\vec{y}$. This can be shown by induction on i, where $\vec{y}$ is allowed to have values outside $[0, h]$ on its first i coordinates. The base case ($i = 0$) we have just seen. For the inductive step, use the fact that $f = f_i$, and f_i is of degree h in direction i. □

We note that $f = f_i$ may hold for all i except one (say, for all $2 \leq i \leq m$), and still be very far from being of degree h. Consider the function $f_{0/1}$ that is 0 whenever $y_1 < n/2$ and 1 whenever $y_1 \geq n/2$. This function is of degree 0 in each of the variables $y_2, \ldots, y_m$, as it does not depend on them. Consider now an arbitrary degree-h function g, and let $f = g + f_{0/1}$. Then $f = f_i$ for $2 \leq i \leq m$. On the other hand, f is far from every degree h multinomial. It has distance roughly $1/2$ from g (which is a degree h multinomial). As every two different degree h multinomials agree on at most a fraction of $mh/|\mathcal{F}|$ of their values, it follows that $\Delta_{DH}(f) \geq 1/2 - mh/|\mathcal{F}|$, which approaches $1/2$ as $|\mathcal{F}|$ grows.

Let $\tau(f)$ denote the fraction of choices of $\vec{y}$ and i on which the basic degree-h test rejects f. We wish to bound $\tau(f)$ from below as a function of $\Delta_{DH}(f)$. The above example shows that the best that we can hope for is a general bound of the form $\tau(f) \geq \Omega(\Delta_{DH}(f)/m)$.

1.1 Analysis of the degree h test

It is indeed true that $\tau(f) \geq \Omega(\Delta_{DH}(f)/m)$ (for sufficently large $|\mathcal{F}|$). This was established in [1]. Here we shall prove a weaker bound, namely, $\tau(f) \geq \Omega(\Delta_{DH}(f)/hm)$, which suffices to establish that $NP \subset PCP(\log n, \text{polylog } n)$. The proof that we present is a straightforward extension of the analysis given for the case $h = 1$ in [2]. We shall indicate how the proof can be strengthened to avoid loosing the $1/h$ factor, without giving full details.

A few words on the cardinality of the finite field $\mathcal{F}$ are in order. Recall that $h \simeq \log n$, $m \simeq \log n / \log\log n$, and that $|\mathcal{F}| \geq hm$ is required for the sum-check protocol. For the proof of $NP \subset PCP(\log n, \text{polylog } n)$ to go through, one needs $|\mathcal{F}|$ to be bounded by a polynomial in h, m (otherwise $O(\log n)$ random bits would not suffice for the verifier in order to sample a random point in $\mathcal{F}^m$). The proof that we give for $\tau(f) \geq \Omega(\Delta_{DH}(f)/hm)$ works when $|\mathcal{F}| \geq h^2 m$. The proof in [1] for $\tau(f) \geq \Omega(\Delta_{DH}(f)/m)$ is claimed to work when $|\mathcal{F}| \geq h^3 m^2$ (though it appears that $|\mathcal{F}| \geq h^2 m^2$ also suffices). Observe that the cardinality of $|\mathcal{F}|$ plays a major role in the size of the PCP witness, and keeping $|\mathcal{F}|$ as small as possible is a major goal in attempts to construct PCPs in which the size of the PCP witness is not much larger than the size of the "traditional" NP witness. For the special case of $m = 2$, it is shown in [3] that $\tau(f) \geq \Omega(\Delta_{DH}(f))$ even when $|\mathcal{F}| = O(h)$. The proof in [3] deteriorates rapidly as m grows, but it turns out that in most PCP constructions (as we will hopefully see in future lectures), the case $m = 2$ is of key importance.

Theorem 2 *Let $|\mathcal{F}| \geq 100h^2m$, and let $f : \mathcal{F}^m \to \mathcal{F}$ be an arbitrary function, $\Delta_{DH}(f) \geq \frac{1}{10}$. Then $\tau(f) \geq \frac{1}{6hm}$.*

Proof: Our proof is composed of two main parts. The first part (Lemma 3) shows that $\tau(f) = \Omega(\min[\Delta_{DH}(f), (1-\Delta_{DH}(f))]/m)$, and is useful for us whenever $\Delta_{DH}(f) \leq 1-1/h$. The second part of the proof (Lemma 5) deals with the case that $\Delta_{DH}(f) > 1-1/h$. It is based on induction on m, and the results of the first part of the proof help the induction step go through.

We now proceed with a detailed proof of Theorem 2.

Lemma 3 *Let $f : \mathcal{F}^m \to \mathcal{F}$ be an arbitrary function. Then $\tau(f) \geq \frac{(1-\Delta_{DH}(f))\Delta_{DH}(f)}{m} - \frac{h}{|\mathcal{F}|}$.*

Proof: Let L be a degree-h function such that $\Delta(f, L) = \Delta_{DH}(f)$. Let G be the indicator function of $L - f$ (i.e. G is 0 where f agrees with L and 1 otherwise). Rather than choose a direction i and one random point c, consider the experiment of choosing a direction i and two random points a and b, both on the same aligned line in direction i. Later we shall choose c to be one of $\{a, b\}$ at random. We say that a pair $\{a, b\}$ is *two colored* if $G(a) \neq G(b)$. Let $\mathcal{E}$ denote the event that $\{a, b\}$ is two colored.

Lower bounding the number of two colored pairs:

The points a and b are each chosen at random with uniform probability from the set $\mathcal{F}^m$. Had they been chosen independently, then $Prob(\mathcal{E}_1)$ would have been exactly $2(1 - \Delta_{DH}(f))\Delta_{DH}(f)$.

However, a and b are not independent, as they are chosen to agree on all their coordinates but one. To quantify the effect of this dependency, we present a two stage processes for choosing a and b, which is equivalent to the actual process used.

1. Select two points p, q indepentently at random from $\mathcal{F}^m$.

2. Select an index i at random, between 1 and m. Let a and b both agree with p on their first $i-1$ coordinates, both agree with q on their last $m-i$ coordinates, and for the i^{th} coordinate, point a agrees with p, whereas point b agrees with q.

Clearly, if $G(p) \neq G(q)$, then there exists a choice of i such that $\mathcal{E}_1$ holds. It follows that $prob(\mathcal{E}_1) \geq \frac{2(1-\Delta_{DH}(f))\Delta_{DH}(f)}{m}$.

Upper bounding the number of good pairs:

Consider an aligned line ℓ in direction i. The total number of pairs $\{a, b\}$ on ℓ is $|\mathcal{F}|(|\mathcal{F}|-1)$. How many such pairs can be two colored and yet both points pass the degree h test in direction i? Call such pairs *good.* To bound the number of good pairs, we use the fact than any two different degree h univariate polynomials agree on at most h points. Let p_1 be the degree h polynomial describing f_i on ℓ, and let p_2 be the degree h polynomial describing L on ℓ. If $p_1 \neq p_2$ as polynomials, then there are at most h points c with $G(c) = 0$ for which $f(c) = p_1(c)$, and the number of good pairs is at most $2h(|\mathcal{F}| - h)$. If $p_1 = p_2$ as polynomials, then there are at most h points c with $G(c) = 0$ for which $f(c) = p_1(c)$, and again the number of good pairs is at most $2h(|\mathcal{F}| - h)$. In both cases, the fraction of good pairs is at most $2h/|\mathcal{F}|$.

Combining the lower bound and the upper bound, the fraction of pairs on which the degree-h test fails is at least $\frac{2(1-\Delta_{DH}(f))\Delta_{DH}(f)}{m} - \frac{2h}{|\mathcal{F}|}$. Selecting one member of the pair at random, the proof of Lemma 3 follows. $\square$

The lower bound on $\tau(f)$ in Lemma 3 improves as $\Delta_{DH}(f)$ grows, up to the point where $\Delta_{DH}(f) = 1/2$. Thereafter, the lower bound on $\tau(f)$ starts to deteriorate, and becomes too weak for Theorem 2 roughly when $\Delta_{DH}(f) > 1 - 1/mh$. Lemma 5 (to follow) is capable of addressing also the case of very large $\Delta_{DH}(f)$, and its prof partly uses the results of Lemma 3. Specifically, it uses the following corollary.

Corollary 4 *Let $|\mathcal{F}| \geq 100h^2m$, and let $f : \mathcal{F}^m \to \mathcal{F}$ be an arbitrary function, $\frac{1}{4(h+1)} \leq \Delta_{DH}(f) \leq 1 - \frac{1}{4(h+1)}$. Then $\tau(f) \geq \frac{1}{5hm}$.*

We note here that the constants chosen in Corollary 4 and elsewhere in this section are to some extent arbitrary, and are given only for concreteness. The computations with these constants work out provided that h is large enough (which we take as a convention here). For very small values of h (e.g., $h = 1$), it may be necessary to replace some of the constants by larger ones.

We are now ready to address the case that $\Delta_{DH}(f)$ is large.

Lemma 5 *Let $|F| \geq 100h^2m$ and let $f : \mathcal{F}^m \to \mathcal{F}$ be an arbitrary function satisfying $\Delta_{DH}(f) \geq \frac{1}{4(h+1)}$. Then $\tau(f) \geq (1 - h/|\mathcal{F}|)^{(m-1)}\frac{1}{5hm}$.*

Proof: The proof is by induction on m. For the base case of the induction ($m = 1$), we need to prove that $\Delta_{DH}(f) > 1/4(h+1)$ implies that $\tau(f) > 1/5h$. When f is a univariate function we have that $\Delta_{DH}(f) \leq \Delta(f, f_1) = \tau(f)$, and the proof follows (when $h \geq 4$).

For the induction step, we prove the statement for m by fixing the first coordinate, and using the induction hypothesis on the other $m - 1$ coordinates. When fixing the value of the first coordinate x_1 to $a \in \mathcal{F}$ we get a subspace $\mathcal{F}_{x_1=a}$. Let us denote by f_a the restriction of f to $\mathcal{F}_{x_1=a}$. Let T_a be the set of all aligned lines in direction x_1 that go through $\mathcal{F}_{x_1=a}$ and let $T'_a \subseteq T_a$ be the set of those x with $x_1 = a$ on which $f(x) \neq f_1(x)$ (or equivalently, the basic degree-h test fails when $i = 1$). Let $\tau_a = |T'_a|/|T_a|$.

Lemma 6 *If along the first coordinate there are $h+1$ distinct values $a_i \in \mathcal{F}$ that satisfy for every $1 \leq i \leq h+1$,*

1. $\Delta_{DH}(f_{a_i}) < \frac{1}{4(h+1)}$ *(where Δ_{DL} here relates to functions on $m-1$ variables)*
2. $\tau_{a_i} < \frac{1}{4(h+1)}$

then $\Delta_{DL}(f) \leq 1/2 + \Delta(f, f_1)$.

Proof: Let us first explain the main point in the proof. If $\Delta(f, f_1)$ is small, then on aligned lines in direction x_1 the function f is described well by univariate polynomials of degree h. We need to show that these polynomials are related in the sense that there is one degree-h m-variate function such that its restriction to a typical line in direction x_1 gives the respective degree-h polynomial. Such a function will be given by Equation (1) below, and the two conditions of the lemma will allow us to show that this function is close to f.

For $1 \leq i \leq h+1$, let L_{a_i} be degree h functions on $\mathcal{F}_{x_1=a_i}$ such that $\Delta(L_{a_i}, f_{a_i}) < 1/4(h+1)$. Let $p_i(x_1)$ be the unique degree h polynomial with $p_i(a_i) = 1$ and $p_i(a_j) = 0$ for all $j \neq i$, $1 \leq j \leq h+1$. Define

$$L(x_1, \ldots, x_m) = \sum_{i=1}^{h+1} p_i(x_1) L_{a_i}(x_2, \ldots, x_m) \tag{1}$$

By definition, $L(x_1, \ldots, x_m)$ is a formal polynomial of degree at most h in each one of its m variables. Observe that by the first condition in Lemma 6, for at least a fraction of $1 - (h+1)/4(h+1) = 3/4$ of the values for $(x_2, \ldots, x_m)$, f agrees simultaneously with all L_{a_i}. Likewise, by the second condition in Lemma 6, for a fraction of at least $3/4$ of the values for $(x_2, \ldots, x_m)$, f_1 agrees simultaneously with all L_{a_i}. Hence on at least $1/2$ of the values for $(x_2, \ldots, x_m)$, f and f_1 simulatneously agree with all L_{a_i}. For every one of the aligned lines determined by these values $(x_2, \ldots, x_m)$, there is a unique degree h univariate polynomial that agrees with the $h+1$ values at $x_1 = a_i$. Hence f_1 and L are represented by the same degree h polynomials along these lines. This shows that $\Delta(f_1, L) \leq 1/2$. By the triangle inequality, $\Delta_{DH}(f) \leq \Delta(f, f_1) + \Delta_{DH}(f_1)$. $\square$

Now we finish the proof of Theorem 2 by a case analysis.

Case 1. $\Delta(f, f_1) \geq 1/4$. Then degree h tests in direction x_1 reject with probability at most $1/4$, showing that $\tau(f) \geq 1/4m$.

Case 2. $\Delta(f, f_1) \le 1/4$, and the conditions of Lemma 6 hold. In this case $\Delta_{DH}(f) \le 3/4$. Moreover, we assumed that $\Delta_{DH}(f) \ge \frac{1}{4(h+1)}$. By Corollary 4, $\tau(f) \ge \frac{1}{5hm}$.

Case 3. $\Delta(f, f_1) \le 1/4$, and the conditions of Lemma 6 do not hold.

Then there is probability $1 - h/|\mathcal{F}|$ of picking $x_1 = b$ with b not satisfying one of the two conditions in Lemma 6. If the condition $\tau_b < 1/4(h+1)$ is the one not satisfied, then with probability $1/m$ the degree h test is performed in direction x_1, and then it rejects with probability $1/4(h+1)$. If the condition $\Delta_{DH}(f_b) < 1/4(h+1)$ is not satisfied, then there is probability $1 - 1/m$ of performing the degree-h test in a direction different then x_1. The induction hypothesis holds on this f_b of dimension $m-1$, and $\tau(f_b) \ge (1 - h/|\mathcal{F}|)^{(m-2)}/5h(m-1)$. Hence

$$\tau(f) \ge (1 - \frac{h}{|\mathcal{F}|}) \min[(1 - \frac{1}{m})\tau(f_b), \frac{1}{4(h+1)}] \ge (1 - \frac{h}{|\mathcal{F}|})^{(m-1)} \frac{1}{5hm}$$

as desired.

$\square$

In order to complete the proof of Theorem 2 we simplify the bound obtained in Lemma 5, using the assumption that $|\mathcal{F}| \ge 30hm$.

$$\tau(f) \ge (1 - \frac{h}{|\mathcal{F}|})^{(m-1)} \frac{1}{5hm} \ge \frac{1}{6hm}$$

$\square$

1.2 Towards tighter analysis

The analysis shows that when $\Delta_{DH}(f) \ge 1/10$, $\tau(f) \ge \Omega(1/hm)$. Here we indicate how the analysis can be modified so as to show $\tau(f) \ge \Omega(1/m)$.

Let us first see where the factor of $1/h$ was lost. For all f with $1/10 \le \Delta_{DH}(f) \le 9/10$, Lemma 3 shows that $\tau(f) \ge \Omega(1/m)$, as desired. So the problem is only with cases that $\Delta_{DH}(f)$ is very large. To analyse them we use Lemma 6 that contains conditions such as $\Delta_{DH}(f_a) < 1/4(h+1)$ and $\tau_a < 1 - 1/4(h+1)$. In the former case we eventually need to apply Lemma 3 with $\Delta_{DH} \simeq 1/h$ (which motivated the formulation of Corollary 4), loosing a factor of h in the bounds (and incidently, also in the size of the finite field $\mathcal{F}$). In the latter case, we need to consider aligned lines in direction $i = 1$ on which only a $1/h$ fraction of the points fail the basic degree-h test. Again, we loose a factor of $1/h$.

To avoid loosing a factor of $1/h$, Lemma 6 needs to be strengthened so as to replace the $1/4(h+1)$ terms by some absolute constants independent of h. To be able to do this, we will need to increase the number of points a_i that we consider. This leads to the following lemma, whose proof will suffice to establish that $\tau(f) \ge \Omega(1/m)$.

Lemma 7 *If along the first coordinate there are $10h$ distinct values $a_i \in \mathcal{F}$ that satisfy for every $1 \leq i \leq 10h$,*

1. $\Delta_{DH}(f_{a_i}) < \frac{1}{10}$,
2. $\tau_{a_i} < \frac{1}{10}$,

then $\Delta_{DL}(f) \leq 1/2 + \Delta(f, f_1)$.

How does one prove Lemma 7? The key to the proof of Lemma 6 was equation (1), which gives a function which is a formal polynomial of degree h. Then, using the fact that many aligned lines in the direction of x_1 had degree-h univariate polynomials that agree with all f_{a_i}, it was shown that the function in equation (1) is not too far from f. For the proof of Lemma 7 we shall also derive a candidate degree-h polynomial, but doing so will be more complicated. The reason for the extra complication is that it might be the case that no aligned line in direction x_1 has a degree-h univariate polynomials that agrees with all respective f_{a_i}. In order to make the proof work, we settle for having many aligned lines in which some degree h polynomial agrees with the vast majority of the f_{a_i}. Then, as a replacement to Equation (1), we use a procedure of Berelekamp and Welsh [4], to be described below. The exact way in which we use it will not be described. (The reader is referred to [1] for this.)

1.3 Error correction

Let $\mathcal{F}$ be a finite field, and let $S = \{(x_i, y_i)\}$ for $1 \leq i \leq n \leq |\mathcal{F}|$ be a set of pairs specifying points $x_i \in \mathcal{F}$ and values $y_i \in F$. We assume that all x_i are distinct. For a parameter t, we are interesting in finding a degree d polynomial $q(x)$ such that for at least t pairs in S, $q(x_i) = y_i$. Observe that if $2t > n + d$, then if such a polynomial q exists then it must be unique, because two different degree d polynomials agree on at most d points.

A brute force approach to finding q when it exists is to pick $d+1$ pairs in S, hope that on them q is correct, and interpolate a polynomial from them. Checking whether this polynomial is the desired q is easy. If the $d+1$ points are chosen at random, the probability of success is $\binom{t}{d+1}/\binom{n}{d+1}$ which is exponentially small in d (with something like t/n as the base of the exponent). The approach of [4] to finding q that we describe now takes time polynomial in d.

Assume that $2t > n + d$ and that q indeed exists. Call a pair $(x_i, y_i) \in S$ an *error point* if $q(x_i) \neq y_i$. We do not know which are the error points, but there are at most $n - t$ of them. Hence there is some nonzero polynomial $e(x)$ of degree $n - t$ (the *error locator polynomial*) that is 0 on the corresponding x_is of all error points. Fixing such an e, let $p(x) = q(x)e(x)$. Then $p(x)$ is a degree $n - t + d$ polynomial.

None of the polynomials p, q, and e is known to us, but we know that for every $(x_i, y_i) \in S$, $p(x_i) = y_i e(x_i)$. This allows us to get a system of n linear equations in which the variables are the coefficients of the polynomials p and e. This system has $(n-t+1)+(n-t+d+1) = 2n-2t+d+2 \leq n+1$ variables. In fact, we can always assume that the leading coefficient of e is 1, leaving only n variables.

By the way we set up the system of equations, it must have a nontrivial solution, namely, the coefficients of two polynomials p' and e'. However, the solution need not be unique, hence p' and e' need not be the original p and e that we had in mind. Nevertheless, we would like to show that if we divide p' by e' as formal polynomial, that they do indeed divide, and that their quotient polynomial q' (satisfying $q'e' = p'$ as formal polynomials) is indeed our desired q.

Consider the polynomial qe'. It is of degree $n-t+d$, which is also the degree of p'. q agrees with t of the pairs in S, so there are t points where $p'(x) = q(x)e'(x)$. As $t > n-t+d$, p' and qe' are identical as polynomials. Hence q can serve as the ratio between p' and e'. It is the unique ratio polynomial, because the ring of polynomials over a finite field is a so called *unique factorization domain*. (Informally, meaning that polynomials behave like integers with respect to factoring. There are *ireducible polynomials* that are the analogues of prime numbers, and every polynomial has a unique factorization into ireducible polynomials, up to multiplication by field elements.)

References

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

[2] U. Feige, S. Goldwasser, L. Lovasz, S. Safra and M. Szegedy. "Interactive Proofs and the Hardness of Approximating Cliques". *Journal of the ACM, Vol. 43, No. 2, March 1996, pp. 268–292.*

[3] A. Polishchuk, D. Spielman. "Nearly-linear Size Holographic Proofs". *Proceedings of 26th ACM Symposium on Theory of Computing*, 194–203, 1994.

[4] L. Welch, E. Berlekamp. "Error correction of algebraic block codes". US Patent Number 4,633,470. (Filed 1986.)