

Algebraic Fingerprinting

Randomized Algorithms Motwani and Raghavan, sections 7.1-7.2
Notes by SCOPI Alfeld
23.9.08

Section 7.1 - *Fingerprinting and Freivalds' Technique*

Multiplying two $n \times n$ matrices A and B can be done in $O(n^3)$ time via the standard matrix multiplication algorithm. The fastest alternative algorithm currently known runs in $O(n^{2.376})$. Verifying, however, that

$$AB = C$$

can be done in $O(n^2)$ time with a randomized algorithm by exploiting the fact that we can multiply a matrix with a vector in $O(n^2)$ time.

Let $r \in \{0, 1\}^n$ be a random vector¹. Let

$$x = Br$$

$$y = Ax = ABr$$

$$z = Cr$$

Clearly

$$(AB = C) \Rightarrow (y = ABr = Cr = z)$$

Obviously, the converse is not true. We will show, however, that

$$P(y = z \mid AB \neq C) \leq \frac{1}{2}$$

Consider the $n \times n$ matrix $D = AB - C$. We note that $P(y = z) = P(Dr = 0)$. Because $AB \neq C$, we know that D is not the zero-matrix. Let the k -th row of D , d^T , be some row with a non-zero entry. Further, let z be the index of some non-zero entry in d^T . The k -th element of Dr is the inner product $d^T r$. Clearly, $P(Dr = 0) \leq P(d^T r = 0)$. We note that $d^T r = 0$ iff

$$r_z d_z = - \sum_{i \neq z} d_i r_i$$

This is to say

$$r_z = \frac{- \sum_{i \neq z} d_i r_i}{d_z}$$

The right hand side of the above equation can be 1, 0, or some other number. Because r_z is either 0 or 1 with equal likelihood, we find the (tight²) upper bound

$$P(y = z) = P(Dr = 0) \leq P(d^T r = 0) \leq \frac{1}{2}$$

By this analysis, we see that the decision to draw elements of r from $\{0, 1\}$ was arbitrary. Any two distinct elements would lead to the same bound. Further, if instead we draw elements of r from some set \mathcal{S} , then

$$P(y = z) \leq \frac{1}{|\mathcal{S}|}$$

¹Each element of r is picked uniformly from the set $\{0, 1\}$

²If D 's first element is 1, and all the rest are 0, for example, then we get $P(r_z = 0) = \frac{1}{2}$ by the construction of r .

Section 7.2 - Verifying Polynomial Identities

We now extend this technique to verifying polynomial multiplication. Given polynomial functions $f_1(x)$, $f_2(x)$ of degree n , and $f_3(x)$ of degree $2n$, we want to verify

$$f_1(x)f_2(x) = f_3(x)$$

Multiplying $f_1(x)$ and $f_2(x)$ takes $O(n\log(n))$ time using the Fast Fourier Transform. We will show that, using a randomized algorithm, we can verify the multiplication in $O(n)$ time.

Let $r \in \mathcal{S}$ be a random number¹. We claim the polynomial multiplication to be valid if

$$f_1(r)f_2(r) = f_3(r)$$

Clearly,

$$\left(f_1(r)f_2(r) \neq f_3(r)\right) \Rightarrow \left(f_1(x)f_2(x) \neq f_3(x)\right)$$

but the converse is not true. In parallel with the previous section, we will show that

$$P\left(\left(f_1(r)f_2(r) = f_3(r)\right) \mid \left(f_1(x)f_2(x) \neq f_3(x)\right)\right) \leq \frac{2n}{|\mathcal{S}|}$$

We let

$$Q(x) = f_1(x)f_2(x) - f_3(x)$$

Because the degree of f_1 and f_2 is n , Q has at most $2n$ distinct roots. Therefore,

$$P\left(Q(r) = 0 \mid Q(x) \neq 0\right) \leq \frac{2n}{|\mathcal{S}|}$$

We can illustrate this pictorially. The graph of $Q(x)$ crosses the x -axis at most $2n$ times, therefore when we select r , the probability that it is one of these zeros is at most $\frac{2n}{|\mathcal{S}|}$. We can also use a deterministic approach of simply picking $2n + 1$ points, and evaluating $Q(x)$ for each. However, the naive approach to this takes $O(n^2)$ time, and even a more advanced technique brings this only to $O(n\log^2(n))$ time (longer than multiplying the original polynomials directly).

In general, this procedure can be used to determine a polynomial identity $f_1(x) = f_2(x)$ by simply letting $Q(x) = f_1(x) - f_2(x)$. Given $f_1(x)$ and $f_2(x)$ explicitly, we can also compare coefficients directly - a deterministic $O(n)$ procedure. This cannot be done, however, when the polynomials are instead given implicitly. One key advantage of this randomized algorithm is that it allows a black-box view of the polynomials.

We now extend this technique to handle the case for multivariate polynomials. First we adopt the book's change of variables in Q . Rather than the degree of the polynomials, n now represents the number of variables. The degree of a term in Q is the sum of the exponents, and the total degree d of Q is the maximum of the degrees of its terms. We now choose a random vector $r \in \mathcal{S}^n$, and want to prove

$$P\left(Q(r_1, \dots, r_n) = 0 \mid Q(x_1, \dots, x_n) \neq 0\right) \leq \frac{d}{|\mathcal{S}|}$$

We prove this by induction on n . Our base case

$$P\left(Q(r_1) = 0 \mid Q(x_1) \neq 0\right) \leq \frac{d}{|\mathcal{S}|}$$

follows from our preceding analysis. We then factor out x_1 (or any other x of our choosing), defining $0 < k \leq d$ to be the greatest exponent of x_1 . This leads to

$$Q(x_1, \dots, x_n) = \sum_{i=0}^k x_1^i Q_i(x_2, \dots, x_n)$$

where $Q_i(x_2, \dots, x_n)$ is the coefficient of x_1^i . We note that $Q_i(x_2, \dots, x_n)$ is a polynomial of $n - 1$ variables, and has degree at most $d - i$.²

¹We draw r from a uniform distribution over \mathcal{S} .

² x_1^i is degree i , and the total degree is i plus the degree of Q_i , which is at most d .

We now consider the polynomial $Q_k(x_2, \dots, x_n)$. The total degree of Q_k is at most $d - k$, and by our choice of k , we know that

$$Q_k \not\equiv 0$$

We now consider the two cases.

Case I: $Q_k(r_2, \dots, r_n) \neq 0$

By our induction hypothesis, we know that $P\left(Q_k(r_2, \dots, r_n) = 0\right) \leq \frac{(d-k)}{|\mathcal{S}|}$.

Case II: $Q_k(r_2, \dots, r_n) = 0$

We define a new, univariate, polynomial $q(x_1) = Q(x_1, r_2, r_3, \dots, r_n)$. By the fact that the coefficient of x_1^k is $Q_k(r_2, \dots, r_n)$, we know that q has degree k , and $q \not\equiv 0$. Because q is univariate, we know that $P(q(r_1) = 0) \leq \frac{k}{|\mathcal{S}|}$

We now know that

$$P\left(Q_k(r_2, \dots, r_n) = 0\right) \leq \frac{d-k}{|\mathcal{S}|}$$

And that

$$P\left(Q(r_1, \dots, r_n) = 0 \mid Q_k(r_2, \dots, r_n) \neq 0\right) \leq \frac{k}{|\mathcal{S}|}$$

Exercise 7.3 in the book shows that for any two events A and B

$$P(A) \leq P(A \mid \bar{B}) + P(B)$$

Therefore

$$P\left(Q(r_1, \dots, r_n) = 0 \mid Q(x_1, \dots, x_n) \not\equiv 0\right) \leq \frac{d-k}{|\mathcal{S}|} + \frac{k}{|\mathcal{S}|} = \frac{d}{|\mathcal{S}|}$$

Exercise 7.3

For events A and B we want to show

$$P(A) \leq P(A | \bar{B}) + P(B)$$

We know that

$$P(A) = P(A \wedge B) + P(A \wedge \bar{B})$$

and, by Bayes' Rule

$$P(A | \bar{B}) = \frac{P(A \wedge \bar{B})}{P(\bar{B})}$$

and clearly,

$$\frac{P(A \wedge \bar{B})}{P(\bar{B})} \geq P(A \wedge \bar{B})$$

Therefore

$$P(A) \leq P(A \wedge B) + \frac{P(A \wedge \bar{B})}{P(\bar{B})} = P(A \wedge B) + P(A | \bar{B}) \leq P(B) + P(A | \bar{B})$$

An example:

Suppose

$$\begin{aligned} Q(x_1, x_2) &= 1 + x_1 + 2x_2 + 3x_1x_2 + 4x_1^2x_2 + 5x_2^2x_1 + 6x_1^2x_2^3 \\ &= x_1^0(1 + 2x_2) + x_1^1(1 + 3x_2 + 5x_2^2) + x_1^2(4x_2 + 6x_2^3) \end{aligned}$$

Here we see that

$$\begin{aligned} n &= 2 \\ d &= 5 \\ k &= 2 \end{aligned}$$

Therefore

$$\begin{aligned} Q_k(x_2) &= Q_2(x_2) = 4x_2 + 6x_2^3 \\ Q_2(r_1) &= 4r_1 + 6r_1^2 \end{aligned}$$

Because $d - k \geq 3$,

$$P\left(Q_2(r_1) = 0\right) \leq \frac{3}{|\mathcal{S}|}$$

We then see that¹

$$q(x_1) = Q(x_1, r_2) = x_1^0(1 + 2r_2) + x_1^1(1 + 3r_2 + 5r_2^2) + x_1^2(4r_2 + 6r_2^3)$$

By the induction base case,

$$P\left(q(r_1) = 0\right) \leq \frac{2}{|\mathcal{S}|}$$

Finally we see that

$$P\left(Q(r_1, r_2) = 0\right) \leq \frac{3}{|\mathcal{S}|} + \frac{2}{|\mathcal{S}|} = \frac{5}{|\mathcal{S}|}$$

¹For example, if $r_2 = 1$, we have $q(x_1) = 3 + 9x_1 + 10x_1^2$.