CPSC 536N: Randomized Algorithms 2014-15 Term 2

Lecture 17

Prof. Nick Harvey *University of British Columbia*

1 Polynomial Identity Testing

In the first lecture we discussed the problem of testing equality of two bitstrings in a distributed setting. We reduced that problem to the problem of testing whether a certain polynomial q is the zero polynomial. That was our first glimpse of polynomial identity testing (PIT).

Let us now define the general PIT problem, while still being a bit vague. Given two multivariate polynomials $p(x_1,\ldots,x_n)$ and $q(x_1,\ldots,x_n)$ over some field $\mathbb{F}$, we would like to decide if they are equal. Note that this is equivalent to deciding if their difference $p-q$ is zero. This equivalent form is often more convenient to think about because it only involves the single polynomial $p-q$. So we could also define PIT to be: given a polynomial $p(x_1,\ldots,x_n)$ over some field $\mathbb{F}$, decide if p is zero.

You might think this problem is rather dull and algebraic, but it actually captures many interesting and natural computational problems. For example, we've already seen that testing equality of strings reduces to PIT (even univariate PIT). We'll see another example in Section 2.

Next we discuss an important way in which our definition of PIT is too vague.

What is a zero polynomial? Our definition of PIT did not say precisely what it means for p to be zero. There are two possible meanings, which lead to two different computational problems. Our desired definition of PIT uses the second meaning.

- The **Evaluates to Zero Everywhere** (EZE) problem. Given a polynomial $p(x_1,\ldots,x_n)$ over $\mathbb{F}$, we must decide whether, for every choice of numbers $y_1,\ldots,y_n \in \mathbb{F}$, the value of $p(y_1,\ldots,y_n)$ is the number 0.
- The **Polynomial Identity Testing** (PIT) problem. Given a polynomial $p(x_1,\ldots,x_n)$, we can write it as a sum over monomials with various coefficients. For example, given $p(x,y,z)=(x+2y)(3y-z)$, we can expand it into a sum of monomials as

  $$p(x,y,z) \;=\; 3xy+6y^2-xz-2yz.$$

  The problem is to decide whether, after expanding p into monomials, are all coefficients of those monomials equal to zero? If so, we say that p is the zero polynomial, or that it is **identically zero**.

It might never have occurred to you that EZE and PIT are different problems. If p is identically zero then definitely it evaluates to zero everywhere. Is the converse true? Over any *infinite* field, like $\mathbb{R}$ or $\mathbb{C}$, the converse is true: a polynomial evaluates to zero at every point if and only if it is identically zero. But over *finite* fields the converse is not true, so EZE and PIT are different problems. For example, the univariate polynomial $p(x)=x^2+x$ over $\mathbb{F}_2$ is not identically zero but $p(0)=0$ and $p(1)=1+1=0$.

It is worth pointing out that trivial approaches do not solve these problems efficiently. The trivial approach for EZE is to exhaustively try every possible choice of numbers $y_1,\ldots,y_n$ and evaluate $p(y_1,\ldots,y_n)$. This requires $|\mathbb{F}|^n$ evaluations. The trivial approach for PIT is to explicitly expand the polynomial into monomials, but if p has degree d then there can be $\binom{n+d}{d}$ such monomials, which is exponential in d.

Let me also clarify what is meant by the degree of a multivariate polynomial (or even monomial). A monomial is any expression of the form $\alpha \cdot \prod_{i=1}^{n} x_i^{\beta_i}$ where $\alpha \in \mathbb{F}$ and $\beta_1, \ldots, \beta_n$ are non-negative integers. The **total degree** of that monomial is $\sum_i \beta_i$. The total degree of a polynomial is defined to be the largest total degree of its monomials.

Our definitions of EZE and PIT are still vague in another important way.

How to represent a polynomial? What does it mean to be *given* a polynomial? One natural definition is to allow $p(x_1, \ldots, x_n)$ to be presented as an explicit algebraic formula involving only the variables $x_1, \ldots, x_n$, any numbers in $\mathbb{F}$, parentheses, and the operations of addition, subtraction and multiplication. For example, $p(x_1, x_2) = x_1 \cdot (x_1 - 2 \cdot x_2)$.

Alternatively, we could allow p to be presented as a "black box", meaning that we do not have an explicit representation of p but our algorithms are allowed to evaluate p by plugging in any desired numbers for $x_1, \ldots, x_n$. This is useful because we might have some weird representation for p that is not an explicit formula. For example, let

$$p(x_1, \ldots, x_n) \;=\; \det \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \ldots & x_2^{n-1} \\ & & & \vdots & \\ 1 & x_n & x_n^2 & \vdots & x_n^{n-1} \end{pmatrix}.$$

Then p is actually a polynomial in $x_1, \ldots, x_n$ of total degree at most $n(n-1)$. We can easily choose numeric values for $x_1, \ldots, x_n$ and evaluate p at that point simply by plugging those numbers into the matrix and computing the determinant numerically. But it might be the case that p does not have a concise representation as an explicit formula. (In this particular example the matrix is a Vandermonde matrix, and the explicit formula $p(x_1, \ldots, x_n) = \prod_{i<j}(x_i - x_j)$ is known.)

For our purposes, it does not matter whether the polynomial is given as an explicit formula or as a black box. We discuss this issue further in Section 1.4.

1.1 Complexity Status

Unfortunately EZE is coNP-hard. There is a simple procedure to encode any 3SAT formula as a polynomial over $\mathbb{F}_2$ such that the polynomial evaluates to zero everywhere if and only if the formula is unsatisfiable.

The complexity status of PIT is much more interesting. We will show that there is a randomized algorithm to decide PIT. However, there is no known deterministic algorithm for deciding PIT. Furthermore, if a deterministic algorithm existed then there would be remarkable consequences in complexity theory.

1.2 The Schwartz-Zippel Lemma

The main tool that we will use today is the Schwartz-Zippel lemma.

Lemma 1 *Let $p(x_1, \ldots, x_n)$ be a polynomial of total degree d. Assume that p is not identically zero. Let $S \subseteq \mathbb{F}$ be any finite set. Then, if we pick $y_1, \ldots, y_n$ independently and uniformly from S,*

$$\Pr[\; p(y_1, \ldots, y_n) = 0 \;] \;\leq\; \frac{d}{|S|}.$$

To help understand the lemma, consider the case of polynomials over $\mathbb{R}$. The theorem says that if we evaluate p at a random point, then we have small probability of seeing a root. Does that mean that polynomials over $\mathbb{R}$ have finitely many roots? No! Consider the polynomial $p(x_1, x_2) = x_1$ which has total degree $d = 1$. It has *infinitely* many roots because setting $x_1 = 0$ and x_2 to anything gives a root. However, if we fix $S = \{0, 1\}$ and randomly choose $x_1, x_2 \in S$ then our probability of seeing a root is exactly $1/2$, which matches the bound given by the Schwartz-Zippel lemma.

It is also worth noting that the conclusion of the theorem does not depend on n.

PROOF: We proceed by induction on n.

The base case is the case $n = 1$, which is the univariate case we discussed in the first lecture. We claimed that any univariate polynomial of degree d has at most d roots. (The reason is that any univariate polynomial of degree d factors uniquely into at most d irreducible polynomials, each of which has at most one root.) So, the probability that y_1 is a root is at most $d/|S|$.

Now we assume the theorem is true for polynomials with $n-1$ variables, and we prove it for those with n variables. The main idea is to obtain polynomials with fewer variables by factoring out the variable x_1 from p. Let k be the largest power of x_1 appearing in any monomial of p. Then

$$p(x_1, \ldots, x_n) \;=\; \sum_{i=0}^{k} x_1^i \cdot q_i(x_2, \ldots, x_n).$$

By our choice of k, the polynomial q_k is not identically zero. Furthermore its total degree is at most $d - k$, so by induction

$$\Pr[\; q_k(y_2, \ldots, y_n) = 0 \;] \;\leq\; \frac{d-k}{|S|}.$$

Define $\mathcal{E}_1$ to be the event "$q_k(y_2, \ldots, y_n) = 0$".

Let us now randomly choose the values of $y_2, \ldots, y_n$ and assume that the event $\mathcal{E}_1$ did not occur. Define $f(x_1)$ to be the univariate polynomial

$$f(x_1) \;=\; \sum_{i=0}^{k} x_1^i \cdot q_i(y_2, \ldots, y_n) \;=\; p(x_1, y_2, \ldots, y_n).$$

Since $\mathcal{E}_1$ did not occur, the coefficient of x_1^k in f is non-zero, so f is not identically zero. By the argument of our base case,

$$\Pr[\, f(y_1) = 0 \;\mid\; \neg\mathcal{E}_1 \,] \;\leq\; \frac{k}{|S|}.$$

Define $\mathcal{E}_2$ to be the event "$f(y_1) = 0$". By definition of f, $\mathcal{E}_2$ is equivalent to "$p(y_1, \ldots, y_n) = 0$".

We're almost done. The goal of the lemma is to bound $\Pr[\,\mathcal{E}_2\,]$, and so far we've bounded $\Pr[\,\mathcal{E}_2 \mid \neg\mathcal{E}_1\,]$ and $\Pr[\,\mathcal{E}_1\,]$. By a few manipulations, we get

$$\begin{aligned}
\Pr[\,\mathcal{E}_2\,] &= \Pr[\,\mathcal{E}_2 \wedge \mathcal{E}_1\,] \;+\; \Pr[\,\mathcal{E}_2 \wedge \neg\mathcal{E}_1\,] \\
&= \Pr[\,\mathcal{E}_1\,] \cdot \Pr[\,\mathcal{E}_2 \mid \mathcal{E}_1\,] \;+\; \Pr[\,\mathcal{E}_2 \mid \neg\mathcal{E}_1\,] \cdot \Pr[\,\neg\mathcal{E}_1\,] \\
&\leq \Pr[\,\mathcal{E}_1\,] \;+\; \Pr[\,\mathcal{E}_2 \mid \neg\mathcal{E}_1\,] \\
&\leq \frac{d-k}{|S|} + \frac{k}{|S|} \;=\; \frac{d}{|S|}
\end{aligned}$$

which finishes the proof. □

1.3 Solving PIT

With the Schwartz-Zippel lemma in hand, we can easily solve PIT. Suppose we are given a polynomial $p(x_1, \ldots, x_n)$ of total degree d. Assume that $d < |\mathbb{F}|$. (For more on this point, see Section 1.4.)

Our algorithm simply picks $y_1, \ldots, y_n$ uniformly and independently from $\mathbb{F}$ (or if $\mathbb{F}$ is infinite, a suitably large finite subset of $\mathbb{F}$). If $p(y_1, \ldots, y_n)$ evaluates to a non-zero value, the algorithm announces that "p is not identically zero". In this case we will never make an error: we have conclusive proof that p is not identically zero. Otherwise, if $p(y_1, \ldots, y_n) = 0$, the algorithm announces that "p is identically zero". In this case the Schwartz-Zippel lemma shows that probability of error is at most $d/|\mathbb{F}| < 1$.

We can decrease the probability of error to any desired level by repeating the algorithm multiple times. This is the amplification by independent trials trick. For example, even if $d = |\mathbb{F}| - 1$, then repeating the algorithm $|\mathbb{F}|$ times reduces the probability of failure to at most

$$\Big(\frac{|\mathbb{F}|-1}{|\mathbb{F}|}\Big)^{|\mathbb{F}|} \;\leq\; \Big(1 - \frac{1}{|\mathbb{F}|}\Big)^{|\mathbb{F}|} \;\leq\; 1/e.$$

1.4 Further Discussion of Complexity Status

Our randomized algorithm above assumed that p had total degree $d < |\mathbb{F}|$. Under that assumption, the Schwartz-Zippel lemma implies that any polynomial that evaluates to zero everywhere must be identically zero. Therefore, under the assumption that $d < |\mathbb{F}|$, the problems EZE and PIT are actually equivalent, so they can both be solved by our randomized algorithm. This also explains why EZE and PIT are equivalent for infinite fields.

It remains to discuss the case $d \geq |\mathbb{F}|$. In this case:

- EZE is NP-hard as mentioned above.
- PIT in the black box model cannot be solved. This is because there are polynomials which are not identically zero, but evaluate to zero everywhere (such as $p(x) = x^2 + x$ over $\mathbb{F}_2$) and these cannot be distinguished from the identically zero polynomial.
- PIT in the explicit formula model is solvable! Since $d \geq |\mathbb{F}|$, the Schwartz-Zippel lemma does not give any useful information about the number of roots. The trick is to use a field extension of $\mathbb{F}$. Any finite field $\mathbb{F}$ can be extended to larger finite field $\mathbb{F}'$ which contains $\mathbb{F}$ as a subfield. So instead of viewing p as a polynomial over $\mathbb{F}$, we view it as a polynomial over a larger field $\mathbb{F}'$ with $d < |\mathbb{F}'|$. This preserves the property of p being identically zero, so we can simply run our randomized algorithm for p over the field $\mathbb{F}'$.

2 Bipartite Matching

We conclude by describing using PIT to solve the Bipartite Matching problem.

Let $G = (U \cup V, E)$ be a bipartite graph, meaning that U and V are disjoint sets of vertices, and every edge in E has exactly one endpoint in U and exactly one endpoint in V. A **matching** in G is a set of edges that share no endpoints. A **perfect matching** in G is a set of edges $M \subseteq E$ such that every vertex is contained in exactly one edge of M.

Polynomial time algorithms are known to decide if G has a perfect matching, and even to construct such a matching. We will give a randomized algorithm to decide if G has a perfect matching, by reducing that problem to PIT.

Let A be the matrix whose rows are indexed by the vertices in U and columns are indexed by the vertices in V. The entries of A are:

$$A_{u,v} \;=\; \begin{cases} x_{u,v} & (uv \in E) \\ 0 & (uv \notin E) \end{cases},$$

where $\{\, x_{u,v} \,:\, uv \in E \,\}$ are distinct variables.

Claim 2 $\det A$ *is identically zero if and only if G has no perfect matching.*

PROOF: By the Leibniz formula for determinants,

$$\det A \;=\; \sum_{\pi} \operatorname{sign}(\pi) \prod_{u \in U} A_{u,\pi(u)},$$

where the sum is over all bijections $\pi : U \to V$ and $\operatorname{sign}(\pi)$ is a function taking values in $\{+1, -1\}$ whose definition is irrelevant for our purposes.

The key observation is that a bijection from U to V is simply a pairing $\{\, (u, \pi(u)) \,:\, u \in U \,\}$ of elements in U and elements in V such that each vertex appears in exactly one pair. This is almost the same as our definition of a perfect matching. If each of those pairs $(u, \pi(u))$ were an edge in E, then that pairing would be exactly a perfect matching in G.

Conveniently, the monomial $\prod_{u \in U} A_{u,\pi(u)}$ tells us exactly when that happens. More precisely, $\prod_{u \in U} A_{u,\pi(u)}$ is a non-zero monomial if and only if π corresponds to a perfect matching in G. Furthermore, since all of these monomials involve distinct sets of variables, there can be no cancellations when we add up the monomials.

Therefore $\prod_{u \in U} A_{u,\pi(u)}$ appears as a monomial in $\det A$ (possibly with a minus sign) if and only if π corresponds to a matching in G. $\square$

This claim yields the following algorithm for testing if a bipartite graph has a perfect matching. Let $\mathbb{F}$ be any field of size at least n^2. We can view $\det A$ as a polynomial over $\mathbb{F}$ since all coefficients of its monomials are $+1$ and -1. Assign every variable $x_{u,v}$ a random value from $\mathbb{F}$. Compute the numeric determinant $\det A$. If the determinant is zero, say "G has no perfect matching". Otherwise, say "G has a perfect matching". Since $\det A$ has degree n the Schwartz-Zippel lemma implies that the failure probability of this algorithm is at most $n/n^2 < 1/n$.