Pattern Matching

In this lecture we start with the problem of pattern matching. Informally, the problem is to find if a given pattern X = $X_1X_2 \ldots X_n$, occurs in a text Y = $Y_1Y_2 \ldots Y_m$, where n $\leq$ m and X,Y$\in \Sigma^*$. Formally, to find is there a $j \leq (m-n)$ such that

$$Y_jY_{j+1} \ldots Y_{j+n-1} = X_1X_2 \ldots X_n$$

We can assume wlog that Σ = {0,1}.

A brute force deterministic algorithm for this problem is :
For $j = 1$ to (n-m+1) do
if Number(Y(j)) = Number(X) then there is a match starting at position j
Update Y(j) to Y(j+1)

Here Number(X) denotes the decimal representation of the binary string X.
Alternatively, we can do Number(Y(j)) $\equiv$ Number(X) mod p where p is a prime. But this can lead to false matches, i.e. Number(Y(j)) $\neq$ Number(X) but Number(Y(j)) = Number(X) mod p. False matches can be detected by explicit verification in O(n) time. The idea is to try to minimize the probability of a false match so that the expected running time is small.

Theorem 10.1 (Chinese Remainder Theorem). *For primes* $p_1, p_2, \ldots p_k$, $x \equiv y$ *mod* p_i $\forall i$ *iff* $(x \equiv y)$ *mod* $(p_1p_2 \ldots p_k)$ *and* $x = \sum c_id_iy_i$ *mod* M *where* $M = (p_1p_2 \ldots p_k)$ *and* $d_i = p_1p_2p_{i-1}p_{i+1} \ldots p_k$.
Also $c_id_i \equiv 1(mod\ p_i)$ *and* $y_i \equiv y$ *mod* p_i.

Randomized Algorithm for Pattern Matching

Choose k primes s.t. $p_1p_2 \ldots p_k$ (=M) $\geq 2^n$.
Hence if Number(Y(j)) $\equiv$ Number(X) mod p_i $\forall$ i, then Number(Y(j)) = Number(X) from the Chinese Remainder Theorem.
Now expand the set of primes to a set of 2k primes p_1, p_2, $\ldots$ p_{2k}. Choose one prime p at random from the set of primes p_1, p_2, $\ldots$ p_{2k} and check if Number(Y(j)) = Number(X) mod p. If true return match else do not.

Lemma 10.2. *The probability of a false match by the above algorithm is less than* $\frac{1}{2}$.

Proof: From the Chinese Remainder Theorem, no more than k-1 primes in the set p_1, p_2, $\ldots$ p_{2k} will satisfy the congruence Number(Y(j)) = Number(X) mod p unless Number(Y(j)) = Number(X). Hence the probability of a false match $\leq \frac{1}{2}$. Note that for that CRT to be applicable, product of any k primes in the set of 2k primes should be $\geq 2^n$. $\square$

We can extend the result to a set of tk primes, and hence the prob. of a false match (at any fixed position) $\leq \frac{1}{t}$. If we choose t $\geq$ m^2, then prob of false match $\leq \frac{1}{m}$. For this we need to have a set of tk primes. Using the Prime Number Theorem, we get

$$\frac{x}{lnx} < \frac{\pi(x)}{tk} < \frac{1.26x}{lnx}$$

$\Rightarrow x \backsim tklog(tk)$, hence size(x) $\backsim$ log t + log k + log(log(tk)) $\backsim$ O(log m).

Verifying Polynomial identities

Suppose we want to test if two polynomials $p_1(x)$ and $p_2(x)$ are identical where the polynomials are not given explicitly. It is easy to see that it is the same as checking: is the polynomial (p_1 - p_2) identically zero. The basic idea is to choose a random value α for the indeterminate x and check if ($p_1(\alpha)$ - $p_2(\alpha)$) = 0. If it is zero, declare $p_1(x)$ and $p_2(x)$ to be equal, otherwise not. The intuition being that since there are only a few zeroes of a polynomial of a given degree, picking a zero at random from a large set is unlikely. This is expressed more formally in the next theorem.

Theorem 10.3 (Schwartz-Zippel Theorem). *Let I be a finite subset of the field F of coefficients and $p(x_1, x_2, \ldots x_n)$ is a multinomial of degree d. If $p(x_1, x_2, \ldots x_n)$ is not identically 0, then the number of roots of p in I^n is bounded by $|I|^{n-1}$ degree($p(x_1, x_2, \ldots x_n)$).*

Proof: Proof is by induction on n. The base case n = 1 involves a polynomial in one variable of degree d, and from algebra a polynomial of degree d cannot have more than d roots. Hence the result follows. Now assume that the induction hypothesis is true for a multivariate polynomial of upto n-1 variables, $n > 1$. Consider a polynomial $P(x_1, x_2, \ldots, x_n)$ of degree n and represent P as $\sum_{i=0}^{k} x_1^i P_i(x_2, x_3, \ldots, x_n)$, where $k \leq d$. Suppose $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ is a root of P. The degree of $P_k(x_2, x_3, \ldots, x_n)$ is atmost d-k, hence using induction hypothesis it can have atmost $|I|^{n-2}$ (d-k) roots. Hence P can have atmost $|I|^{n-1}$ (d-k) roots in the case $P_k(x_2, x_3, \ldots, x_n) = 0$. Suppose that, $P_k(x_2, x_3, \ldots, x_n) \neq 0$. We can consider P as a univariate polynomial of degree k. Hence it can have atmost $|I|^{n-1}$(k) roots in this case for different choices of $(x_2, x_3, \ldots, x_n)$. Therefore P can have atmost $|I|^{n-1}$ degree($p(x_1, x_2, \ldots x_n)$) roots. □

Lemma 10.4. *If $p(x_1, x_2, \ldots x_n)$ is a non-zero polynomial of degree d with coefficients from field F and I is a subset of F, then the probability that p evaluates to zero on a random element (i_1, i_2,...,i_n) $\in I^n$ is atmost $\frac{d}{|I|}$.*

Proof: Proof follows directly from the theorem above. □

Applications of polynomial zero testing

Given three n $\times$ n matrices A, B and C over a field F, consider the problem of verifying if AB = C. Using the best known matrix multiplication algorithm this problem can be solved in O(n$^{2.376}$) time. Using the polynomial identity testing algorithm above, we can develop a O(n^2) randomized algorithm for this problem which has a bounded error probability. It is easy to see that AB(x) and C(x) where x is a vector can be computed in O(n^2). The intuitive idea here is that if AB = C then AB(x) = C(x) and if AB $\neq$ C, then we want to bound the error probability that we get AB(x) = C(x). Let D = AB - C. The algorithm chooses a random vector x ϵ I^n; where each component of x is choosen independently and uniformly at random from I, where I $\subseteq$ F. We compute AB(x) and

C(x) and output true if AB(x) = C(x), else returns false. It is easy to see that if AB(x) $\neq$ C(x) then AB $\neq$ C. The next theorem gives a bound on the probability that the algorithm outputs true and AB $\neq$ C.

Theorem 10.5. *Let A, B and C be n $\times$ matrices over F such that AB $\neq$ C. Then for a vector x choosen uniformly at random from some subset I of F, Pr[AB(x) = C(x)] $\leq \frac{1}{|I|}$.*

Proof: Let D = AB - C and D is not the all zeroes matrix. We can consider D as a multinomial of degree one and using the Schwartz-Zippel Theorem, Pr[D(x) = 0] $\leq \frac{1}{|I|}$. □

Perfect Matchings

A matching M of a undirected graph G(V,E) is a subset of edges such that each vertex occurs atmost once in M and a perfect matching is a matching of size exactly $\frac{|V|}{2}$. Consider the problem of checking if a given unweighted graph has a perfect matching. Associate with each edge ij of the graph G a distinct variable x_{ij}. The Tutte Matrix for a unweighted graph G is a $|V| \times |V|$ matrix T(G), such that $t_{ij} = x_{ij}$ if $ij = e \in E$ and $i < j$, t_{ij} = -x_{ij} if $ij = e \in E$ and $i > j$, and 0 otherwise.

Note that the Tutte Matrix has $\frac{n^2-n}{2}$ indeterminates and is skew symmetric. We denote the determinant of A by $|A|$.

Theorem 10.6 (Tutte's Theorem). *$|T(G)|$ is not identically zero iff G has a perfect matching, here $|T(G)|$ is a polynomial in the variables x_{ij}.*

Theorem 10.7 (Extension to Tutte's Theorem (Lovasz)). *$T(G)$ has rank 2m if the maximum matching has size m.*

The determinant of a matrix A is given by

$$Det(A) = \sum_{\sigma \epsilon \pi_n} sign(\sigma) value(\sigma)$$

Here σ is a permutation and the sign of σ is + or - depending upon if it is a even or odd permutation respectively. π_n is the symmetric group of permutations of size n.
A permutation can be written as the product of cycles (orbits). For example consider the permutation (3 2 4 1). It can be expressed as the product of two cycles (1 3 4)(2). A cycle is odd if it has odd number of elements and vice versa.
An ordered cycle $(c_1, c_2, c_3, \ldots, c_k)$ has an associated graph.

Observation: If the value of a permutation is non-zero and it has no odd-cycles then we can form a matching by including every alternate edge of the trails.

Lemma 10.8. *G has a perfect matching iff there exists a permutation σ s.t. val(σ) $\neq$ 0 and σ does not have an odd cycle.*

Proof (of Tutte's Theorem): We can now prove Tutte's theorem using the lemma stated above. Suppose G has a perfect matching say M. Consider the term (-1)$^{\frac{n}{2}}$ $\sum_i x_{i,M(i)}^2$. Choose $x_{i,M(i)} = 1$ and 0 for the other, then det(T(G)) $\neq$ 0.

Now we prove the converse, i.e. given that det(T(G)) $\neq 0$, to show that there exists a perfect matching. We shall prove this by contradiction. Suppose det(T(G)) $\neq 0$ and there exists no perfect matching.
This implies that all non-zero terms σ contain an odd cycle (from the lemma). Hence if σ contains an odd cycle say $(x_{1,i_1}, x_{2,i_2}, \ldots x_{2k+1,i_{2k+1}})$ then we can get another permutation by reversing the trail. Since it is a skew-symmetric matrix, there exists another σ' s.t. val(σ') = val(σ) and sign(σ') = - sign(σ). These terms can cancel out and hence det(T(G)) = 0, which is a contradiction to the assumption. □