Fourier Transforms & Theoretical Computer Science **Spring 1999**

Lecture 4: February 19

Lecturer: Umesh Vazirani *Scribe: Chris Umans*

Today's topics:

- lower bounds for constant depth circuits (Razborov/Smolensky)
- a learning algorithm for constant depth circuits (Hastad [Has86] and Linial/Mansour/Nisan [LMN89])

4.1 Lower bounds for constant depth circuits

Definition 4.1 *A* constant depth circuit *is a circuit whose input nodes are labeled with $x_1, x_2, \ldots x_n$ and $\overline{x_1}, \overline{x_2}, \ldots \overline{x_n}$, with a constant number of levels of gates, each of which is an AND gate or an OR gate with unbounded fan-in.*

This computational model was first studied in the early 80's by Furst, Saxe and Sipser [FSS84], and Ajtai. We will establish a lower bound for the parity function $x_1 \oplus x_2 \oplus \cdots \oplus x_n$, which can also be viewed as $f(x) = x_1 + x_2 + \cdots + x_n \pmod 2$ if $x_i \in \{0,1\}$ for all i. The parity function is easy to compute for log-depth circuits; a simple recursive construction yields a log-depth parity circuit of size $O(n)$. For constant depth circuits computing parity, we will establish the following lower bound:

Theorem 4.2 *Depth d constant depth circuits computing parity on n variables require at least $2^{\Omega(n^{\frac{1}{2d}})}$ gates.*

The proof has two major parts:

1. We show that any function $f(x_1, x_2, \ldots x_n)$ computed by a constant depth circuit can be *closely approximated* by a low-degree polynomial $p(x_1, x_2, \ldots x_n)$ over $GF(3)$. Here "low-degree" will mean degree $\sqrt{n}$, and it will be sufficient for p to agree with f on 3/4 of the possible inputs.
2. Using Fourier analysis, we show that parity cannot be approximated by such a low degree polynomial. (Notice that to represent parity *exactly* with a polynomial, we need a degree n polynomial, since $\text{parity}(x_1, x_2, \ldots x_n) = \prod_i x_i$, where "true" is represented by -1, and "false" is represented by 1).

4.1.1 Approximating constant-depth circuits with polynomials

Suppose we have a circuit of depth d computing the function $f(x_1, x_2, \ldots x_n)$ using M gates. Our strategy is the following: we replace each gate in turn by a low-degree (2ℓ-degree) polynomial, arguing at each step that the new circuit agrees with the old circuit on all but a small fraction ($1/2^\ell$) of the inputs. In the end we have a *polynomial* of degree $(2\ell)^d$ that agrees with the original *circuit* on all but a $M/2^\ell$ fraction of the inputs.

Lemma 4.3 *There exists a polynomial $g(y_1, y_2, \ldots y_m)$ over $GF(3)$ of degree 2ℓ that approximates an OR gate with inputs $y_1, y_2, \ldots y_m$ correctly on all but a $1/2^\ell$ fraction of the inputs to the circuit.*

Proof: In this section true is represented by 1 and false is represented by 0. As a first attempt, take the approximating polynomial to be a random linear combination of the inputs to the OR gate:

$$c_1 y_1 + c_2 y_2 + \cdots + c_m y_m$$

We are working over $GF(3)$, so this linear combination may take on a value other than 0 or 1. To correct this, we square it, and then by Fermat's Little Theorem, the result is always 0 or 1. Notice that this approximation is always correct if the OR gate should be outputing 0. The probability that this approximating polynomial errs (by outputing 0 when some y_i is 1) is:

$$\Pr_{\vec{c}}[(c_1 y_1 + c_2 y_2 + \cdots + c_m y_m)^2 = 0 | \bigvee_i y_i = 1] = \frac{1}{3}.$$

To see that this is true, imagine picking c_i last. Since $y_i = 1$, there is exactly one value (out of three field elements) that will make the polynomial output 0.

We now reduce this error by combining ℓ of these random linear combinations. The final polynomial approximating the OR gate is:

$$g(y_1, y_2, \ldots y_m) = 1 - \prod_{j=1}^{\ell} [1 - (c_{1j} y_1 + c_{2j} y_2 + \cdots + c_{mj} y_m)^2].$$

It is an easy extension of our analysis above to see that:

$$\Pr_{\vec{c}}[g(y_1, y_2, \ldots y_m) = 0 | \bigvee_i y_i = 1] \leq (1/3)^\ell < (1/2)^\ell.$$

The key observation is that we have constructed a polynomial whose degree depends on ℓ, instead of the fan-in (m) of the OR gate. ■

Theorem 4.4 *Given an M-gate circuit of depth d computing the function $f(x_1, x_2, \ldots x_n)$, we can construct a polynomial $p(x_1, x_2, \ldots x_n)$ of degree at most $\sqrt{n}$ that errs on only a $M(2^{\frac{1}{2} n^{1/2d}})$ fraction of the inputs.*

Proof: The above lemma describes how to approximate an OR gate; we can approximate an AND gate using the polynomial $1 - g(1 - y_1, 1 - y_2, \ldots 1 - y_m)$. Substituting the appropriate "local" approximating polynomial for each gate in the circuit gives a polynomial $p(x_1, x_2, \ldots x_n)$ that approximates the whole function f. The final polynomial p has degree at most $(2\ell)^d$. Now, we analyze the probablility that this polynomial errs:

$$\Pr_{\vec{x}} \Pr_{\vec{c}}[p(x_1, x_2 \ldots x_n) \neq f(x_1, x_2, \ldots x_n)] \leq \frac{M}{2^\ell}$$

By interchanging the order of the probabilities, we can conclude that there *exists* a choice of $\vec{c}$ such that:

$$\Pr_{\vec{x}}[p(x_1, x_2 \ldots x_n) \neq f(x_1, x_2, \ldots x_n) = 1] \leq \frac{M}{2^\ell}.$$

Notice that the extra level of randomization (the c's) was crucial to overcome the possibility that the induced input distribution at any particular gate might be very skewed.

We'd like the approximating polynomial to have degree $\sqrt{n}$. So, we choose $\ell = 1/2n^{1/2d}$, and we get a $\sqrt{n}$ degree polynomial with error at most $M/(2^{\frac{1}{2} n^{1/2d}})$. ■

4.1.2 Impossibility of approximating parity with a low-degree polynomial

In past lectures, we used the Fourier transform on functions $f : \mathbb{Z}_2^n \to \mathbb{C}$. In this section, we are concerned with the space of all functions $f : \mathbb{Z}_2^n \to \mathbb{Z}_3$. It turns out that the Fourier basis here is still the parity basis that is familiar from the other setting. In the normal basis, we represent true by 1 and false by 0; in the parity basis, we represent true by -1 and false by 1. Therefore, we can change from the normal basis to the Fourier basis by substituting $y + 1$ for the variable x (and from the Fourier basis to the normal basis by substituting $x - 1$ for y). Observe that these are linear substitutions so the degree of the polynomial does not change.

The key observation in this section is that parity is "complete" for the space of all functions $f : \mathbb{Z}_2^n \to \mathbb{C}$ in the following sense:

Lemma 4.5 *Let $D \subseteq \mathbb{Z}_2^n$. If parity can be computed by a degree d polynomial $p : D \to \mathbb{Z}_3$, then* all *functions $f : D \to \mathbb{Z}_3$ can be computed by $n/2 + d$ degree polynomials.*

Proof: Suppose we have such a polynomial p. The basis functions in the Fourier basis are $\chi_S = \prod_{i \in S} y_i$. Any function $f : D \to \mathbb{Z}_3$ can be written as:

$$\begin{aligned} f(y_1, y_2, \ldots y_n) &= \sum_S C_S \chi_S \\ &= \sum_{S:|S| \le n/2} C_S \chi_S + \sum_{S:|S|>n/2} C_S \chi_S \\ &= \sum_{S:|S| \le n/2} C_S \chi_S + \sum_{S:|S|>n/2} C_S \chi_{\overline{S}} p(y_1, y_2, \ldots y_n) \end{aligned}$$

Notice that in this last equation all of the χ_S have degree at most $n/2$, so the total degree is at most $n/2 + d$. ■

Now we can prove Theorem 4.2. Let p be a $\sqrt{n}$ degree polynomial approximating parity obtained from an M gate depth d circuit as described in the previous section. Take D to be the portion of $\mathbb{Z}_2^n$ on which p is correct. The above lemma implies that any function $f : D \to \mathbb{Z}_3$ can be written as a degree $n/2 + \sqrt{n}$-degree polynomial.

The number of monomials of degree at most $n/2 + \sqrt{n}$ is

$$\sum_{j=0}^{n/2+\sqrt{n}} \binom{n}{j} \le (1-c)2^n \text{ for some constant } 0 < c \le 1/2.$$

All higher degree monomials, of which there are at least $c2^n$, must lie *outside* the region D, because otherwise we could compute them with a degree $n/2 = \sqrt{n}$ polynomial, a contradiction. Therefore, the fraction of inputs on which p errs must be at least c. So we have:

$$M/(2^{\frac{1}{2} n^{1/2d}}) \ge c$$

which implies that $M \ge c(2^{\frac{1}{2} n^{1/2d}})$. Hence parity requires exponential size constant depth circuits.

4.2 Learning constant-depth circuits

Suppose we are given a function f that can be computed by a constant depth, polynomial size circuit. We want to learn this function by observing m samples $\{(\vec{x_i}, f(\vec{x_i}))\}_{i=1}^m$ selected uniformly at random. In other words, we want to formulate a hypothesis $h(x_1, x_2, \ldots, x_n)$ for which $\Pr_{\vec{x}}[f(\vec{x}) = h(\vec{x})] \geq 1 - \epsilon$ with high probability $(1-\delta)$. Using Fourier methods, Linial, Mansur, and Nisan show how to achieve ϵ error after looking at $O(n^{poly \log n})$ samples. The hypothesis h is evaluatable in quasi-polynomial time in n, $1/\epsilon$ and $1/\delta$.

The key lemma states that the higher Fourier coefficients of such a function f are small, so we can get away with estimating only the lowest $O(n^{poly \log n})$ coefficients, and taking the higher coefficients to be zero.

In the analysis, we need to consider our functions to be real-valued functions from $\mathbb{Z}_2^n$ to the interval $[-1, 1]$ (a Boolean function in this setting maps TRUE to 1 and FALSE to -1). The reason we cannot use a finite field as in the previous section is that we need to use Parseval's identity, which requires an *ordered field.* Recall Parseval's identity:

Theorem 4.6 (Parseval's Identity) $E(f^2) = \sum_S (\hat{f}(S))^2$.

In order to use Fourier analysis for this problem, we bound the accuracy of the hypothesis h as follows (which will eventually allow us to apply Parseval):

Claim 4.7 $\Pr_{\vec{x}}[f(\vec{x}) \neq \mathit{sign}(h(\vec{x}))] \leq E[(f-h)^2]$

Proof: For each $\vec{x}$, either (1) $f(\vec{x}) = \text{sign}(h(\vec{x}))$, in which case the contribution to the LHS is 0 and the contribution to the RHS is ≥ 0; or (2) $f(\vec{x}) \neq \text{sign}(h(\vec{x}))$, in which case the contribution to the LHS is 1 and the contribution to the RHS is ≥ 1, (since $f(\vec{x}) \in \{1, -1\}$). ■

Our task now is to show that the high Fourier coefficients are small, and that we can approximate the lower Fourier coefficients well.

4.2.1 Bounding the high Fourier coefficients

The main lemma (which we prove in a later section) is this:

Lemma 4.8 (Main [LMN89]) *If f is computable by a depth d size M circuit, then:*

$$\sum_{S:|S|>t} \hat{f}(S)^2 \leq M \cdot 2^{-\frac{1}{4}t^{1/(d+3)}}.$$

When $M = n^{O(1)}$ and $t = O(\log^{\Theta(d)} n/\epsilon)$, this implies $\sum_{S:|S|>t} \hat{f}(S)^2 \leq \epsilon/2$.

4.2.2 Estimating the lower Fourier coefficients

We estimate the lower Fourier coefficients from the m samples $\{(\vec{x_i}, f(\vec{x_i}))\}_{i=1}^m$ as follows:

$$\alpha_S = \frac{\sum_{i=1}^m f(\vec{x_i})\chi_S(\vec{x_i})}{m}.$$

Notice that α_S is an unbiased estimator of $\hat{f}(S)$, so by a standard application of Chernoff bounds, we get that:

$$\Pr[|\alpha_S - \hat{f}(S)| \leq \sqrt{\frac{\epsilon}{2n^t}} \text{ for all } S : |S| \leq t] \geq 1 - \delta.$$

4.2.3 Analysis of the hypothesis h

We take our hypothesis to be sign(h), where h is the the function whose Fourier coefficients are given by α_S for $S : |S| \leq t$, and zero otherwise. Using claim 4.7 and Parseval's identity, we have:

$$\begin{aligned}
\Pr_{\vec{x}}[f(\vec{x}) \neq \text{sign}(h\vec{x})] \leq E[(f-h)^2] &\leq \sum_S (\hat{f}(S) - \hat{h}(S))^2 \\
&\leq \sum_{S:|S|\leq t} (\hat{f}(S) - \alpha_S)^2 + \sum_{S:|S|>t} (\hat{f}(S) - \hat{h}(S))^2 \\
&= \sum_{S:|S|\leq t} (\hat{f}(S) - \alpha_S)^2 + \sum_{S:|S|>t} (\hat{f}(S))^2 \\
&\leq \epsilon/2 + \epsilon/2 = \epsilon
\end{aligned}$$

To evaluate h, we need to compute $\hat{f}(S)$ for all S for which $|S| \leq t$, which given our choice of t, takes time quasi-polynomial in n, ϵ, and δ.

4.2.4 Proof of main lemma

The main tool in proving the lemma is Hastad's Switching lemma. We consider a *random restriction* in which we assign some fraction of the inputs a constant (0 or 1), and assign the rest of the inputs "*", which means that they remain unchanged. Specficially, $\Pr[x_i = *] = p$, $\Pr[x_i = 0] = 1 - p/2$, and $\Pr[x_i = 1] = 1 - p/2$ for each i. The idea is that a large fan-in AND or OR gate should collapse under such a restriction.

Lemma 4.9 (Hastad Switching Lemma [Has86]) *If function f is computable by a depth d size M circuit. Then under a random restriction ρ with $p = 1/(20k)^d$,*

$$\Pr_\rho[f_\rho \textit{ cannot be evaluated by a decision tree of depth } k] \leq M2^{-2k}$$

Hastad's lemma tells us that with high probability, f_ρ has no high Fourier coefficients. This is because any function f that can be evaluated by a depth k decision tree, must have $\hat{f}(S) = 0$ for $|S| > k$, since such a decision tree can be evaluated by a polynomial of degree at most k. We now argue that hitting f with a random restriction cannot alter the Fourier coefficients by much, because all we are really doing is restricting the following sum:

$$\sum_{\vec{x}} f(\vec{x})\chi_S(\vec{x}) = \hat{f}(S)$$

to just *part* of the sum. Therefore we can conclude that the high Fourier coefficients of f cannot have been large, because they are zero after the restriction.

To prove this simple observation requires viewing a random restriction in a slightly different way than usual. We use the following notation: if $S, R \subseteq \{1 \ldots n\}$, then $f_{\overline{S} \leftarrow R}$ is the restriction obtained by assigning 1 to variables in $\overline{S} \cap R$ and 0 to variables in $\overline{S} - R$. We will think of choosing the random restriction ρ as two steps. First, we choose a set S randomly such that each variable appears in S independently with probability p. Set S is the set of variables that are left alive by the restriction. Second, we choose a set R at random which determines the 0–1 assignment to the variables in $\overline{S}$.

We now state two lemmata without proof:

Lemma 4.10 *Let S be a subset of $\{1 \ldots n\}$. For every $B \subseteq S$, we have:*

$$\frac{1}{2^{|\overline{S}|}} \sum_{R \subseteq \overline{S}} (\hat{f}_{\overline{S} \leftarrow R}(B))^2 = \sum C \subseteq S(\hat{f}(B \cup C))^2.$$

Lemma 4.10 is used to prove:

Lemma 4.11 *Fix $S \subseteq \{1 \dots n\}$. For R chosen uniformly at random, we have:*

$$\sum_{A:|A\cap S|>k} (\hat{f}(A))^2 \leq \Pr_R[f_{\overline{S}\leftarrow R} \textit{ does not have a decision tree of depth } k].$$

Finally, we sketch the proof of the main lemma. Our goal is to bound the sum of squares of the higher Fourier coefficients. Let $S, R \subseteq \{1 \dots n\}$ define a random restriction ρ that leaves each variable alive with probability p as described above. We first note that:

$$\sum_{A:|A|>t} (\hat{f}(A))^2 \leq 2E_S\left[\sum_{A:|A\cap S|\geq pt/2} \hat{f}(A)^2\right].$$

This holds for large enough p (specifically, $pt > 8$), because then the probability that $|A \cap S|$ is large enough to contribute to the RHS is at least $1/2$. Therefore each A contributes $(\hat{f}(A))^2$ to the RHS for at least half of the sets S.

We use lemma 4.11 to bound the RHS above by:

$$2E_S\left[\sum_{A:|A\cap S|\geq pt/2} \hat{f}(A)^2\right] \leq 2E_S\left[\Pr_R[f_{\overline{S}\leftarrow R} \text{ does not have a decision tree of depth } pt/2]\right].$$

Let $u = \sqrt{(pt/2)}$. Hastad's Switching Lemma gives us the following bound:

$$\begin{aligned}2E_S\left[\sum_{A:|A\cap S|\geq pt/2} \hat{f}(A)^2\right] &\leq 2E_S\left[\Pr_R[f_{\overline{S}\leftarrow R} \text{ does not have a decision tree of depth } pt/2]\right] \\ &= 2\Pr[f_\rho \text{ cannot be evaluated by a decision tree of depth } pt/2] \leq 2\frac{M}{2^u}.\end{aligned}$$

For the appropriate choice of u, (which determines p), the lemma follows.

References

[FSS84] M. Furst, J.B. Saxe, M. Sipser. Parity, circuits, and the polynomial-time hierarchy. Mathematical Systems Theory, vol.17(1), April 1984. p.13-27.

[Has86] J. Hastad. Computational limitations for small depth circuits. MIT Press, 1986. Ph.D. thesis.

[LMN89] N. Linial, Y. Mansour, N. Nisan. Constant depth circuits, Fourier transform, and learnability. 30th Annual Symposium on Foundations of Computer Science (FOCS 89). p.574-9.