CS 395T Computational Complexity of Machine Learning

Lecture 15: March 10, 2005 Lecturer: Adam Klivans

Scribe: Andy Mills

Learning Constant Depth Circuits

15.1 Intuition

In this lecture, we will investigate the Fourier coeffiecients of AC^0 circuits and show they are concentrated on the lower degree terms. Thus we can use the low degree algorithm to learn them if we can get input/output examples from the uniform distribution. This result is due to [Linial, Mansour, Nisan 1993]. We will proceed in a top-down manner.

AC^0 circuits are those circuits with polynomial size and constant depth as a function of the size of their input. Note that polynomial size DNF and CNF formulas are both depth two AC^0 circuits, so AC^0 is a very reasonable generalization of what we've done earlier.

The intuition is that [Hastad 1989] first showed that $PARITY$ cannot be computed by AC^0 circuits, and in fact circuits of size roughly $2^{n^{1/depth}}$ cannot even approximate it well. Informally, this implies that AC^0 circuits might have very small concentration on higher degree Fourier coefficients. The key technical machinery defined in Hastad's paper is something called a random restriction. A random restriction takes a boolean function, sets a random subset of its input variables to randomly either 0 or 1, and results in possible simplification to a circuit computing that function. Interestingly, Hastad showed that AC^0 circuits simplify drastically under random restrictions. We will use this property to show that the Fourier spectrum of an AC^0 circuit with size s, depth d, and n input variables is concentrated on terms with degree $\leq \log^{O(d)} s/\varepsilon$, where ε is an accuracy parameter we'll define later. This implies we can use the low degree algorithm to get a learning algorithm that runs in time $n^{\log^{O(d)} s/\varepsilon}$.

15.2 Details

Let's get specific. First, define a random restriction ρ with parameter $p \leq 1$ as something which takes a boolean function f and applies the following operation: for each input variable of the function, leave it alone with probability p, set it to 0 with probability $\frac{1-p}{2}$, and set it to 1 with probability $\frac{1-p}{2}$ to produce a new function called f_ρ. A set A will be called "live" if all the variables in it have not been set to values under ρ.

[Linial, Mansour, Nisan 1993] show that $p \leq \frac{1}{10^d l^{d-1}} \Rightarrow Pr_\rho[Largest-Degree-Of-A-Nonzero-Fourier-Coefficient(Some-AC^0-Circuit_\rho) \geq l] \leq \frac{s}{2^l}$. This isn't too difficult to show from Hastad's result, but it's a little off topic, so we elided it in lecture. Essentially, if you take any AC^0 circuit and take a random restriction that restricts enough variables, very likely you will get a function whose nonzero Fourier coefficients have low maximum degree.

In this lecture, we will prove that $\sum_{|A|>t}\widehat{f}^2(A) \le 2Pr_\rho[Largest - Degree - Of - A - Nonzero - Fourier - Coefficient(f_\rho) \ge \frac{tp}{2}]$ for any f that is computable by an AC^0 circuit. This implies if we let $p = \frac{1}{10t^{\frac{d-1}{d}}}$ and $l = \frac{t^{1/d}}{10}$, we can use the result from the last paragraph to prove that $\sum_{|A|>t}\widehat{f}^2(A) \le s2^{-\frac{t^{1/d}}{10}}$. Setting $t = 10\log^d s/\varepsilon$ for arbitrary ε, we get $\sum_{|A|>10\log^d s/\varepsilon}\widehat{f}^2(A) \le \varepsilon$.

To obtain this result, we will prove three intermediate results:

Lemma 1: $E_\rho[\sum_{|live(A)|\ge tp/2}\widehat{f}^2_\rho(A)] \le Pr_\rho[\exists A(|live(A)| \ge tp/2 \bigwedge \widehat{f}_\rho(A) \ne 0)]$. Observe that the last quantity is definitionally equivalent to the probability that the largest degree of a nonzero Fourier coefficient of f_ρ is greater than or equal to $tp/2$.

Lemma 2: $E_S[\sum_{|A\bigcap S|\ge tp/2}\widehat{f}^2(A)] = E_\rho[\sum_{|live(A)|\ge tp/2}\widehat{f}^2_\rho(A)]$ where the expectation in the first quantity is taken over sets S chosen by randomly including each element of $[1...n]$ with probability p.

Lemma 3: $\sum_{|A|\ge t}\widehat{f}^2(A) \le 2E_S[\sum_{|A\bigcap S|\ge tp/2}\widehat{f}^2(A)]$.

These three lemmas directly imply the result.

15.3 Proof of Lemmas

Proof: [Lemma 1] Create an indicator variable I_ρ which is 1 iff under ρ, there exists a set A such that $|live_\rho(A)| \ge tp/2$ and $\widehat{f}_\rho(A) \ne 0$. So the right side of the equation is $Pr_\rho[I_\rho = 1] = E_\rho[I_\rho]$. Now let's try to bound I_ρ. Consider a ρ for which $I_\rho = 1$. Then $I_\rho = 1 = \sum_A \widehat{f}^2_\rho(A) \ge \sum_{|live(A)|\ge tp/2}\widehat{f}^2_\rho(A)$. For a ρ with $I_\rho = 0$, $I_\rho = 0 = \sum_{|live(A)|\ge tp/2}\widehat{f}^2_\rho(A)$. Therefore, in both cases, $I_\rho \ge \sum_{|live(A)|\ge tp/2}\widehat{f}^2_\rho(A)$. Taking expectation over ρ, the lemma is proved. ■

Proof: [Lemma 2] Fix S as some subset of $[1...n]$. Let A be a dummy variable for summation over the subsets of S, let B be a dummy variable for summation over the subsets of the compliment of S, S^c, and let C be a dummy variable for summation over the subsets of $[1...n]$. Let k be a parameter we'll set later.

To refresh our memories from a previous lecture, remember the definition $f_A(x) = \sum_B \widehat{f}(A\bigcup B)\chi_{restricted-to-S}(x) = E_y[f(xy)\chi_A(y)]$. By squaring and taking expectation over x, this implies $\sum_B \widehat{f}^2(A\bigcup B) = E_x[f_A^2(x)] = E_x[(E_y[f(xy)\chi_A(y)])^2]$. By the completeness relation, the last quantity equals $E_x[\widehat{f}^2_{S^c\leftarrow x}(A)]$. Summing over all choices of A whose intersection with S is at least k, we have $\sum_{|A\bigcap S|\ge k}\sum_B \widehat{f}^2(A\bigcup B) = \sum_{|A\bigcap S|\ge k}E_x[\widehat{f}^2_{S^c\leftarrow x}(A)]$. An equivalent expression is $\sum_{|C\bigcap S|\ge k}\widehat{f}^2(C) = \sum_{|C\bigcap S|\ge k}E_x[\widehat{f}^2_{S^c\leftarrow x}(C)]$. Taking the expectation over all choices of S (where each element of $[1...n]$ is chosen for inclusion with probability p), and recognizing that expectation over S as the live set and x is precisely expectation over ρ, we have $E_S[\sum_{|C\bigcap S|\ge k}\widehat{f}^2(C)] = E_\rho[\sum_{|live(C)|\ge k}\widehat{f}^2_\rho(C)]$. If we take $k = tp/2$, the lemma is proved.

■

Proof: [Lemma 3] For a set A with $|A| > t$, the expected value of $|A\bigcap S|$ is pt, so intuitively the probability that the intersection has size greater than $pt/2$ should be large. In fact, Chernoff's bound gives that this probability is at least $1 - e^{-tp/8}$. If we assume $tp > 8$, then this quantity is at least $1 - 1/e > 1/2$. Therefore, every set A is included in the right hand side for more than half of the choices of S, and the lemma is proved. ■

References

[Hastad 1989] J. Hastad. Almost Optimal Lower Bounds for Small Depth Circuits. Advances in Computing Research 5 (1989) 143-170. (Also appeared in STOC 1986.)

[Linial, Mansour, Nisan 1993] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant Depth Circuits, Fourier Transform, and Learnability. Journal of the Association for Computing Machinery, Vol 40, No 3, July 1993. Preliminary version first appearead in 1989.