COMS 6998-2 Advanced Topics in Computational Learning Theory **Spring 2005**

Lecture 9 : March 23, 2005

Lecturer: Rocco A. Servedio *Scribe: Anshul Kundaje and Andrew Wan*

1 Overview

In this lecture we finish the algorithm for learning half-spaces under the uniform distribution. We also show that, with very little effort, the learning result can be extended to include intersections or even arbitrary boolean functions of a constant number of halfspaces. We then continue to exploit Fourier Analysis techniques to give a quasipolynomial time algorithm for learning constant depth polynomial sized boolean circuits.

2 Halfspaces Have Low Fourier Concentration

We continue the proof of the bound on the Fourier concentration of half-spaces. Recall that last time we presented the low degree algorithm which can learn functions f in time $n^{O(d)}$ if f satisfies $\sum_{|S|>d} \hat{f}(S)^2 \leq \varepsilon$. We also defined the noise sensitivity as

$$ns_\varepsilon(f) = \Pr_{x,\text{noise}}[f(x) \neq f(N_\varepsilon(x))],$$

and saw that the noise sensitivity of the parity function on k variables is $\frac{1}{2} - \frac{1}{2}(1-2\varepsilon)^k$. Using the Fourier representation of boolean functions, we can conclude from this that

$$ns_\varepsilon(f) = \frac{1}{2} - \frac{1}{2}\sum_S (1-2\varepsilon)^{|S|} \cdot \hat{f}(S)^2. \tag{1}$$

Consequently we have that if $\varepsilon < 1/2, ns_\varepsilon(f) < 1/2$ for all f. We also have that the parity function on n variables, PAR_n, has the maximum noise sensitivy over all boolean functions f. This is because the Fourier coefficient of $\hat{f}(S)$ for $S = [n]$ is 1, so the negative term in $ns_\varepsilon(\text{PAR}_n)$ is minimized.

In order to bound the Fourier concentration of half-spaces, we first prove the following claim which relates the weight of high Fourier coefficients to noise sensitivity. Informally, the claim states that if a function f has a lot of weight on its higher Fourier coefficients, then f must have high noise sensitivity.

Claim 1 *For any boolean function* $f : \{-1,1\}^n \to \{-1,1\}$*, for any* $0 < \gamma < 1/2$*,*

$$\sum_{|S|\geq 1/\gamma} \hat{f}(S)^2 \leq 2.32 ns_\gamma(f)$$

Proof: By Equation 1,

$$\begin{aligned}
2ns_\gamma(f) &= 1 - \sum_S (1-2\gamma)^{|S|} \cdot \hat{f}(S)^2 \\
&= \sum_S \hat{f}(S)^2 - \sum_S (1-2\gamma)^{|S|} \hat{f}(S)^2 \qquad \text{by Parseval's Identity} \\
&= \sum_S \left[1-(1-2\gamma)^{|S|}\right] \hat{f}(S)^2 \geq \sum_{S\geq 1/\gamma} \left[1-(1-2\gamma)^{|S|}\right] \hat{f}(S)^2 \\
&\geq \sum_{S\geq 1/\gamma} \left[1-(1-2\gamma)^{1/\gamma}\right] \hat{f}(S)^2 = \left[1-(1-2\gamma)^{1/\gamma}\right] \sum_{S\geq 1/\gamma} \hat{f}(S)^2 \\
&\geq (1-\frac{1}{e^2}) \sum_{S\geq 1/\gamma} \hat{f}(S)^2
\end{aligned}$$

Dividing both sides of the inequality by $(1-\frac{1}{e^2})$ gives us the claim. ■

The claim gives us this corollary:

Corollary 1 *Given a boolean function* $f : \{-1,1\}^n \to \{-1,1\}$*, let* $\beta : [0,\frac{1}{2}] \to [0,\frac{1}{2}]$ *be such that* $ns_\gamma(f) \leq \beta(\gamma)$*. Then* $\sum_{|S|\geq m} \hat{f}(S)^2 \leq \varepsilon$ *for* $m = \frac{1}{\beta^{-1}(\varepsilon/2.32)}$.

Proof:

$$\begin{aligned}
\sum_{|S|\geq m} \hat{f}(S)^2 &= \sum_{|S|\geq \frac{1}{\beta^{-1}(\varepsilon/2.32)}} \hat{f}(S)^2 \\
&\leq 2.32 ns_{\beta^{-1}(\varepsilon/2.32)}(f) \leq 2.32\beta(\beta^{-1}(\varepsilon/2.32)) = 2.32\frac{\varepsilon}{2.32} = \varepsilon
\end{aligned}$$

■

The following fact allows us to choose a bound on the weight of the high Fourier coefficients for half-spaces:

Fact 1 *For any half-space* f*,* $ns_\varepsilon(f) \leq 8.8\sqrt{\varepsilon}$.

In the interest of time, we skip the justification of this fact. Taking $\beta(\gamma) = 8.8\sqrt{\gamma}$ and $\beta^{-1}(x) = (x/8.8)^2$, so that $\frac{1}{\beta^{-1}(\varepsilon/2.32)} = \left(\frac{2.32(8.8)}{\varepsilon}\right)^2 \leq \frac{441}{\varepsilon^2}$, we have the following theorem:

Theorem 1 *For any half-space f,*

$$\sum_{|S|\geq\frac{441}{\varepsilon^2}} \hat{f}(S)^2 \leq \varepsilon.$$

Using the low degree algrotihm, we can obtain a $n^{O\left(\frac{1}{\varepsilon^2}\right)}$ algorithm for learning half-spaces(recall that the previous result for learning half-spaces by the Winnow algorithm has a much better exponent than $\frac{442}{\varepsilon^2}$). Still, this technique easily extends to yield unnprecedented results for arbitrary functions on half-spaces.

Theorem 2 *Let $h_1 \dots h_k$ be half-spaces, and $f = g(h_1,\dots,h_k)$ for some boolean $g : \{-1,1\}^k \to \{-1,1\}$. Then $ns_\varepsilon(f) \leq 8.8k\sqrt{\varepsilon}$.*

Proof: For $1 \leq i \leq k$, we have that $ns_\varepsilon(h_i) \leq 8.8\sqrt{\varepsilon}$. We also have that

$$\begin{aligned} ns_\varepsilon(f) = \Pr_{x,\text{noise}}[f(x) \neq f(N_\varepsilon(x))] &\leq \Pr_{x,\text{noise}}[\exists_i \text{ s.t. } h_i(x) \neq h_i(N_\varepsilon(x))] \\ &\leq \sum_{1\leq i\leq k} ns_\varepsilon(h_i) \leq 8.8k\sqrt{\varepsilon} \end{aligned}$$

where the first inequality holds by a union bound. ■

Now to bound the weight of the high degree Fourier coefficients of any boolean function on k half-spaces, we take $\beta(\gamma) = 8.8k\sqrt{\gamma}$ and $\beta^{-1}(x) = \left(\frac{x}{8.8k}\right)^2$ so that $\frac{1}{\beta^{-1}(\varepsilon/2.32)} = \left(\frac{2.32(8.8k)}{\varepsilon}\right)^2 = \frac{441k^2}{\varepsilon^2}$, and we get the following:

Theorem 3 *Let $h_1 \dots h_k$ be half-spaces, and $f = g(h_1,\dots,h_k)$ for some boolean $g : \{-1,1\}^k \to \{-1,1\}$. Then*

$$\sum_{|S|\geq\frac{441k^2}{\varepsilon^2}} \hat{f}(S)^2 \leq \varepsilon.$$

Thus we can learn any such f in $n^{O\left(\frac{k^2}{\varepsilon^2}\right)}$ time. Note that for *arbitrary* functions on half-spaces, this upper bound is not too far from the lower bound on the low degree algorithm for such functions, since the low degree algorithm needs n^k time to learn f if g is the parity function.

3 Learning Constant Depth Poly Size Circuits

Definition 1 *A boolean circuit C is a directed acyclic graph. There is one output node, and each of the rest of the nodes is a gate $\vee, \wedge, \neg, x_1, \ldots, x_n, 0, 1$, with $\vee$ and $\wedge$ gates having arbitrary fanin. The circuit computes a boolean function in the obvious way. The* ***size*** *of a circuit C is the number of $\vee$ and $\wedge$ gates. The* ***depth*** *of C is the length of the longest path from any input node to the output node.*

The rest of this lecture is devoted to proving the following theorem due to Linial, Mansour, and Nisan.

Theorem 4 (LMN) *Let f be computed by a size M depth D circuit. Then for sufficiently large t,*

$$\sum_{|S|\geq t} \hat{f}(S)^2 \leq 2M2^{-\frac{t^{1/D}}{20}}$$

Taking $t = (20 \log \frac{2M}{\varepsilon})^D$, we have that $2^{-\frac{t^{1/D}}{20}} \leq \varepsilon$. We can then apply the low degree algorithm to learn any f computed by a size M depth D circuit in time $n^{O((\log \frac{M}{t})^D)}$.

The proof of Theorem 4 is quite involved and we will not complete it until the next lecture. We will achieve the result via two main Lemmas, the first of which relies on an important result due to Hastad.

3.1 The Switching Lemma

Definition 2 *A* ***restriction*** *ρ is a mapping of $x_1, \ldots, x_n$ to $\{0, 1, *\}$. The result of applying a restriction ρ to a string $x \in \{0,1\}^n$ is the string $x|_\rho$ whose i'th bit is set to $\rho(x_i)$ if $\rho(x_i) \in \{0,1\}$ and x_i if $\rho(x_i) = *$. Given a boolean function $f : \{0,1\}^n \rightarrow \{0,1\}$ and restriction ρ, f_ρ denotes the function f under restriction ρ, defined by $f_\rho(x) = f(x|_\rho)$. A* ***random restriction*** *with parameter p is chosen by setting each x_i independently to $*$ with probability p, 0 with probability $(1-p)/2$, and 1 with probability $(1-p)/2$.*

Definition 3 *The decision tree depth of a function g, denoted DT-depth(g), is the depth of the shallowest decision tree computing g.*

We will use Hastad's switching lemma, which we state without proof.

Theorem 5 (Hastad's Switching Lemma) *Let f be a d-DNF and ρ be a random restriction with parameter p. Then*

$$\Pr_{\rho}[DT\text{-}depth(f_\rho) \geq s] \leq (5pd)^s.$$

Using this result we can prove the following lemma which we will use in the proof of Theorem 4:

Lemma 1 *Let f be a size M depth D circuit. Fix s, and let ρ be a random restriction with parameter $p \leq \frac{1}{10^d s^{D-1}}$. Then*

$$\Pr_{\rho}[DT\text{-}depth(f_\rho) \geq s] \leq M \cdot 2^{-s}.$$

Proof: Take $p = \frac{1}{10^D s^{D-1}}$. We will view ρ as taking place in a series of D stages. In the first stage we apply a random restriction with parameter $p_1 = 1/10$, and in the following $D-1$ stages we apply a random restriction with parameter $p = 1/(10s)$. In the end we are left with a restriction with parameter $p = \frac{1}{10^D s^{D-1}}$. We use the first stage to try to obtain a circuit whose bottom level gates have fanin at most s. At each subsequent stage we hope that the resulting circuit can be collapsed to a smaller depth circuit, so that after all restrictions have been applied, we are left with a circuit of the desired depth. We will show that each of the M gates of C contributes at most 2^{-s} to the failure probability of this process.

After the first stage, we claim that for any bottom-level gate the probability that the gate has fanin greater than s is at most 2^{-s}. We examine two cases before the restriction: either the gate has fanin F greater than $2s$ or the gate has fanin F at most $2s$. In the first case, the gate reduces to a constant if one of its inputs is set to 0 (if the gate is a $\wedge$) or 1 (if the gate is a $\vee$). Thus the probability that the gate is not reduced to a constant is at most $(.55)^{2s}$. In the second case, we see that the gate will have fanin greater than s only if at least s vars of f are set to $*$. This probability is at most

$$\sum_{i=s}^{F} \binom{F}{i} (.1)^i \leq (.1)^s \sum_{i=s}^{F} \binom{F}{i} \leq 2^F (.1)^s \leq 4^s (.1)^s \leq (4/10)^s.$$

Now, assuming that the first restriction did not 'fail', the bottom-level gates have fanin at most s. Consider any gate at distance 2 from the inputs. WLOG, we may assume by alternation that this is a $\vee$ gate. This $\vee$ is the root of an $s-DNF$, since each of its children is a $\wedge$ with at most s children. By the switching lemma, applying a random restriction with parameter $p = 1/(10s)$ gives us a decision tree of depth at most s with failure probability at most $(5ps)^s = 2^{-s}$. We've seen that such a decision tree has an equivalent $s-CNF$. Doing this for each gate at distance 2 from the inputs, we can collapse the resulting circuit a level by merging the resulting root of each $s-CNF$ with its parent, which by alternation is also a $\wedge$ gate(see Figure 3.1). Now we can repeat this for $d-2$ stages(by duality, we can use

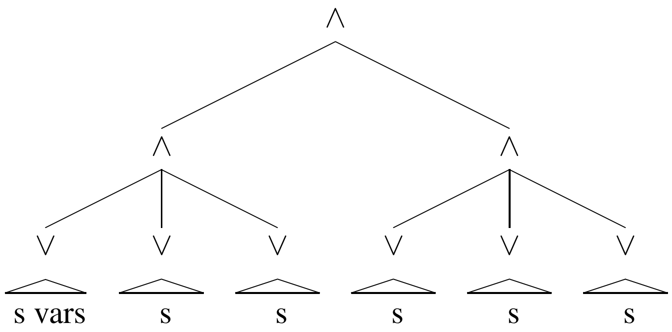

Figure 1: Two Three-Clause s-CNFs collapsing to their parent

the switching lemma to get a small depth decision tree for $s-CNFs$ also) to get a depth $d-(d-2)=2$ circuit that is equivalent to an $s-DNF$, where each gate fails to produce the desired result with probability at most 2^{-s}. We apply the last restiction with $p = 1/(10s)$ to get a decision tree of depth at most s. Since we will obtain a decision tree of this depth unless some gate fails to collapse as desired, and each of M gates fails with probability at most 2^{-s}, the probability of failure is at most $M \cdot 2^{-s}$. ■

We will need yet another lemma to prove Theorem 4.

Lemma 2 *Let f be any boolean function and ρ be a random restriction with parameter p. Then for any t such that $pt \geq 8$,*

$$\sum_{|S|>t} \hat{f}(S)^2 \leq 2 \Pr_{\rho}\left[\textit{DT-depth}(f_\rho) \geq \frac{pt}{2}\right].$$

Modulo Lemmas 1 and 2, the proof of Theorem 4 is easy. Let ρ be a random restriction with parameter $p = \frac{1}{10t^{\frac{d-1}{d}}}$ and assume $pt = \frac{t^{1/d}}{10}$ is at least 8. If we let $s = \frac{pt}{2} = \frac{t^{1/d}}{20}$, then

$$p = \frac{1}{10t^{\frac{d-1}{d}}} < \frac{2^{d-1}}{10t^{\frac{d-1}{d}}} = \frac{20^{d-1}}{10^d t^{\frac{d-1}{d}}} = \frac{1}{10^d s^{d-1}}$$

where the last equality holds because $s^{d-1} = \left(\frac{t^{1/d}}{20}\right)^{d-1}$. Since $p < \frac{1}{10t^{\frac{d-1}{d}}}$, we may apply Lemma 1 to get that

$$\Pr_{\rho}\left[\text{DT-depth}(f_\rho) \geq \frac{pt}{2}\right] \leq M \cdot 2^{-\frac{t^{1/d}}{20}}.$$

Theorem 4 follows from this combined with the inequality in Lemma 2.

3.2 Proof of Lemma 2

Given any decision tree g, if DT-depth$(g) \leq k$, then $\hat{g}(S) = 0$ for any $|S| > k$. This is because we can view g as a sum of "path functions" over k variables, each having zero Fourier coefficients for any $|S| > k$. Since none of these have non-zero Fourier coefficients for any $|S| > k$, neither will g.

Given ρ and $A \subseteq \{x_1, \ldots, x_n\}$, let $live(A) = |\{x_i \in A : \rho(x_i) = *\}|$. If there exists A such that $live(A) \geq \frac{pt}{2}$ and $\hat{f}_\rho(A) \neq 0$, then DT-depth(f_ρ) must be at least $\frac{pt}{2}$. Thus we have

$$\Pr_\rho\left[\exists A : live(A) \geq \frac{pt}{2} \text{ and } \hat{f}_\rho(A) \neq 0\right] \leq \Pr_\rho\left[\text{DT-depth}(f_\rho) \geq \frac{pt}{2}\right]$$

To prove Lemma 2, it suffices to show

$$\sum_{|S|>t} \hat{f}(S)^2 \leq 2 \Pr_\rho\left[\exists A : live(A) \geq \frac{pt}{2} \text{ and } \hat{f}_\rho(A) \neq 0\right],$$

which we do using three facts. Given p and t such that $pt \geq 8$ and $S \subseteq [n]$ chosen by taking each $i \in [n]$ to be in S with probability p, we have the following:

Fact 2

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

Fact 3

$$E_S\left[\sum_{A:|A\cap S|>(pt)/2} \hat{f}(A)^2\right] = E_\rho\left[\sum_{A:live(A)\geq(pt)/2} \hat{f}(A)^2\right],$$

and finally, we have that

Fact 4

$$E_\rho\left[\sum_{A:live(A)\geq(pt)/2} \hat{f}(A)^2\right] \leq \Pr_\rho\left[\exists A : live(A) \geq \frac{pt}{2} \textit{ and } \hat{f}_\rho(A) \neq 0\right]$$

We defer the proofs of Facts 3 and 4 until the next lecture.

Proof: (of Fact 2) Fix any A with $|A| > t$. We have that $E_S[|A \cap S|] > pt$, so by Chernoff bounds, $\Pr_S[|A \cap S| \geq (pt)/2] > 1/2$ if $pt \geq 8$. The contribution of each such A to $E_S\left[\sum_{A:|A \cap S|>(pt)/2} \hat{f}(A)^2\right]$ is exactly $\hat{f}(A)^2 \Pr_S[|A \cap S| \geq (pt)/2]$ which is at least $\frac{1}{2}\hat{f}(S)^2$, so that $E_S\left[\sum_{A:|A \cap S|>(pt)/2} \hat{f}(A)^2\right] \geq \sum_{A:|A|>t} \frac{1}{2}\hat{f}(S)^2$. The claim follows. ■