Lecture 9: Learning DNF, AC0, Juntas

1 Learning DNF in Almost Polynomial Time

From previous lectures, we have learned that if a function f is ϵ-concentrated on some collection $\mathcal{S}$, then we can learn the function using membership queries in $\text{poly}(|\mathcal{S}|, 1/\epsilon)\text{poly}(n)\log(1/\delta)$ time. In the last lecture, we showed that a DNF of width w is ϵ-concentrated on a set of size $n^{O(\frac{w}{\epsilon})}$, and concluded that width-w DNFs are learnable in time $n^{O(\frac{w}{\epsilon})}$.

Today, we shall improve this bound, by showing that a DNF of width w is ϵ-concentrated on a collection of size $w^{O(w\log\frac{1}{\epsilon})}$. We shall hence conclude that $\text{poly}(n)$-size DNFs are learnable in almost polynomial time.

Recall that in the last lecture we introduced Håstad's Switching Lemma, and we showed that DNFs of width w are ϵ-concentrated on degrees up to $O(w\log\frac{1}{\epsilon})$.

Theorem 1.1 *(Håstad's Switching Lemma) Let f be computable by a width-w DNF, If $(\mathbf{I}, \mathbf{X})$ is a random restriction with $*$-probability ρ, then $\forall d \in \mathbb{N}$,*

$$\Pr_{\mathbf{I},\mathbf{X}}[DT\text{-}depth(f_{\mathbf{X}\to\bar{\mathbf{I}}}) > d] \leq (5\rho w)^d$$

Theorem 1.2 *If f is a width-w DNF, then*

$$\sum_{|U|\geq O(w\log\frac{1}{\epsilon})} \widehat{f}(U)^2 \leq \epsilon$$

To show that a DNF of width w is ϵ-concentrated on a collection of size $w^{O(w\log\frac{1}{\epsilon})}$, we also need the following theorem:

Theorem 1.3 *If f is a width-w DNF, then*

$$\sum_{U}\left(\frac{1}{20w}\right)^{|U|}|\widehat{f}(U)| \leq 2$$

Proof: Let $(\mathbf{I}, \mathbf{X})$ be a random restriction with $*$-probability $\frac{1}{20w}$. After this restriction, the DNF becomes a $O(1)$-depth decision tree with high probability. Due to Håstad's Switching Lemma, we

have the following:

$$\mathop{\mathbf{E}}_{\mathbf{I},\mathbf{X}}\left[\|f_{\mathbf{X}\to\bar{\mathbf{I}}}\|_1\right] = \sum_{d=0}^{n} \Pr_{\mathbf{I},\mathbf{X}}\left[\text{DT-depth}(f_{\mathbf{X}\to\bar{\mathbf{I}}}) = d\right] \cdot \mathop{\mathbf{E}}_{\mathbf{I},\mathbf{X}}\left[\|f_{\mathbf{X}\to\bar{\mathbf{I}}}\|_1 \middle| \text{DT-depth}(f_{\mathbf{X}\to\bar{\mathbf{I}}}) = d\right]$$

$$\leq \sum_{d=0}^{n} \left(5 \cdot \frac{1}{20w} \cdot w\right)^d \cdot 2^d \qquad \text{(Håstad's Switching Lemma, DT of size s has } L_1 \text{ Fourier norm} \leq s)$$

$$= \sum_{d=0}^{n} \frac{1}{2^d} \leq 2$$

In addition, we have

$$\begin{aligned} 2 \geq \mathop{\mathbf{E}}_{\mathbf{I},\mathbf{X}}\left[\|f_{\mathbf{X}\to\bar{\mathbf{I}}}\|_1\right] &= \mathop{\mathbf{E}}_{\mathbf{I}}\mathop{\mathbf{E}}_{\mathbf{X}}\sum_{S\subseteq\mathbf{I}}\left|\widehat{f_{\mathbf{X}\to\bar{\mathbf{I}}}}(S)\right| \\ &= \mathop{\mathbf{E}}_{\mathbf{I}}\sum_{S\subseteq\mathbf{I}}\mathop{\mathbf{E}}_{\mathbf{X}}\left[\left|\mathop{\mathbf{E}}_{y\in\{-1,1\}^{\mathbf{I}}}[f_{\mathbf{X}\to\bar{\mathbf{I}}}(y)\mathbf{y}_S]\right|\right] \\ &> \mathop{\mathbf{E}}_{\mathbf{I}}\sum_{S\subseteq\mathbf{I}}\left|\mathop{\mathbf{E}}_{\mathbf{X}}\mathop{\mathbf{E}}_{y}[f(y,\mathbf{X})\mathbf{y}_S]\right| = \mathop{\mathbf{E}}_{\mathbf{I}}\sum_{S\subseteq\mathbf{I}}\left|\widehat{f}(S)\right| \\ &= \sum_{U\subseteq[n]}\left|\widehat{f}(U)\right| \cdot \Pr_{\mathbf{I}}[U\subseteq\mathbf{I}] = \sum_{U\subseteq[n]}\left|\widehat{f}(U)\right| \cdot \left(\frac{1}{20w}\right)^{|U|} \end{aligned}$$

□

Corollary 1.4 *If f is a width-w DNF, then*

$$\sum_{|U|\leq O(w\log\frac{1}{\epsilon})}\left|\widehat{f}(U)\right| \leq w^{O(w\log\frac{1}{\epsilon})}$$

Proof:

$$\begin{aligned} 2 &\geq \sum_{U\subseteq[n]}\left|\widehat{f}(U)\right| \cdot \left(\frac{1}{20w}\right)^{|U|} \geq \sum_{|U|\leq O(w\log\frac{1}{\epsilon})}\left|\widehat{f}(U)\right| \cdot \left(\frac{1}{20w}\right)^{|U|} \\ &\geq \left(\frac{1}{20w}\right)^{O(w\log\frac{1}{\epsilon})} \sum_{|U|\leq O(w\log\frac{1}{\epsilon})}\left|\widehat{f}(U)\right| \end{aligned}$$

□

Corollary 1.5 *If f is a width-w DNF, it's ϵ-concentrated on a collection of size $w^{O(w\log\frac{1}{\epsilon})}$.*

Lecture 9: Learning DNF, AC0, Juntas

Proof: Define $\mathcal{S} = \left\{U : |U| \leq O(w \log \frac{1}{\epsilon}), \left|\widehat{f}(U)\right| \geq \frac{\epsilon}{w^{O(w \log \frac{1}{\epsilon})}}\right\}$. By Parseval, we get that $|\mathcal{S}| \leq w^{O(w \log \frac{1}{\epsilon})}$. We now show that $\mathcal{S}$ is ϵ-concentrated on $\mathcal{S}$. By Theorem 1.2, we know that

$$\sum_{\substack{U \notin \mathcal{S} \\ |U| \geq O(w \log \frac{1}{\epsilon})}} \widehat{f}(U)^2 \leq \epsilon$$

By Corollary 1.4, we have

$$\sum_{\substack{U \notin \mathcal{S} \\ |U| \leq O(w \log \frac{1}{\epsilon})}} \widehat{f}(U)^2 \leq \sum_{\substack{U \notin \mathcal{S} \\ |U| \leq O(w \log \frac{1}{\epsilon})}} \left|\widehat{f}(U)\right| \cdot \max \left|\widehat{f}(U)\right| \leq w^{O(w \log \frac{1}{\epsilon})} \cdot \frac{\epsilon}{w^{O(w \log \frac{1}{\epsilon})}} = \epsilon$$

Therefore, f is 2ϵ-concentrated on $\mathcal{S}$. □

Corollary 1.6 *$\mathrm{poly}(n)$-size DNFs are ϵ-concentrated on collections of size*

$$\left(\log \frac{n}{\epsilon}\right)^{O(\log \frac{n}{\epsilon} \log \frac{1}{\epsilon})} = \left(\frac{n}{\epsilon}\right)^{O(\log \log \frac{n}{\epsilon} \cdot \log \frac{1}{\epsilon})}$$

And if $\epsilon = \Theta(1)$, then the above is $n^{O(\log \log n)}$. Note that this uses the fact that size-n DNF formulas are ϵ-close to a width-$\log(\frac{n}{\epsilon})$ DNF.

An open research problem is the following question: Are $\mathrm{poly}(n)$-size DNFs ϵ-concentrated on a collection of size $\mathrm{poly}(n)$, assuming that $\epsilon = \Theta(1)$.

2 Learning AC^0

We will now study how to learn polynomial-size, constant-depth circuits, AC^0.

Consider circuits with unbounded fan-in AND, OR, and NOT gates. The size of the circuit is defined as the number of AND/OR gates. Observe the following fact:

Fact 2.1 *Let d denote the depth of the circuit. At the expense of a factor of d in size, these circuits can be taken to be "layered". Here "layered" means that eacy layer consists of the same type of gates, either AND or OR, and adjacent layers contain the opposite gates.*

In a layered circuit, the number of layers is the depth of the circuit, and define the *bottom fan-in* of a layered circuit to be the maximum fan-in at the lowest layer (i.e. closest to the input layer).

Theorem 2.2 (LMN.) *Let f be computable by a size $\leq s$, depth $\leq D$, and bottom fan-in $\leq w$ circuit. Then*

$$\sum_{|U| \geq (10w)^D} \widehat{f}(U)^2 \leq O(s \cdot 2^{-w})$$

Lecture 9: Learning DNF, AC0, Juntas

Before we show a proof of the LMN theorem, let us first look at some corollaries implied by this theorem.

Corollary 2.3 *If f has a size s, depth D circuit, then*

$$\sum_{|U|\geq[O(\log\frac{s}{\epsilon})]^D} \widehat{f}(U)^2 \leq \epsilon$$

Proof: Notice that such an f is ϵ-close to a similar circuit with bottom fan-in $\leq \log(\frac{s}{\epsilon})$. □

Corollary 2.4 *AC^0, i.e., the class of poly-size, constant-depth circuits, are learnable from random examples in time $n^{\text{poly}(\log(\frac{n}{\epsilon}))}$, where n denotes the size of the circuit.*

Proof: According to Corollary 2.3, AC^0 circuits are ϵ-concentrated on a collection of size $n^{\text{poly}(\log(\frac{n}{\epsilon}))}$. □

Corollary 2.5 *If f has a size-s, depth-D circuit, then*

$$\mathbb{I}(f) \leq [O(\log s)]^D$$

Remark 2.6 *Due to Håstad, the above bound can be improved to $[O(\log s)]^{D-1}$.*

Proof:(sketch.) Define $\widehat{F}(\tau) = \sum_{|U|>\tau} \widehat{f}(U)^2$. Recall that

$$\begin{aligned}\mathbb{I}(f) = \sum_U |U|\widehat{f}(U)^2 = \sum_{r=1}^{n} \widehat{F}(r) &= \sum_{r=1}^{O(\log s)^D} \widehat{F}(r) + \sum_{r=O(\log s)^D}^{n} \widehat{F}(r) \\ &= \sum_{|U|\leq O(\log s)^D} |U|\widehat{f}(U)^2 + \widehat{F}(O(\log s)^D)\cdot O(\log s)^D + \sum_{r=O(\log s)^D}^{n} \widehat{F}(r) \\ &\leq O(\log s)^D + \sum_{r=O(\log s)^D}^{n} \widehat{F}(r)\end{aligned}$$

It remains to show that $\sum_{r=O(\log s)^D}^{n} \widehat{F}(r) \leq O(\log s)^D$. By Corollary 2.3,

$$\widehat{F}(\tau) = \sum_{|U|>\tau} \widehat{f}(U)^2 \leq s\cdot 2^{-\Omega(\tau^{1/D})}$$

Using this fact plus some manipulation, it is not hard to show that $\sum_{r=O(\log s)^D}^{n} \widehat{F}(r) \leq O(\log s)^D$. □

Using this fact, we can derive the following two corollaries:

Lecture 9: Learning DNF, AC0, Juntas

Corollary 2.7 *Parity $\notin AC^0$, Majority $\notin AC^0$.*

Proof: $\mathbb{I}(\text{Parity}) = n$, $\mathbb{I}(\text{Majority}) = \Theta(\sqrt{n})$. □

Definition 2.8 (PRFG.) *A function $f : \{-1,1\}^m \times \{-1,1\}^n \to \{-1,1\}$ is a Pseudo-Random Function Generator (PRFG), if for all Probablistic Polynomial Time (P.P.T) algorithm A with access to a function,*

$$\left| \underset{\substack{S\subseteq\{-1,1\}^m,\\ A\text{'s random bits}}}{\mathbf{Pr}} [A(f(S,\cdot)) = YES] - \underset{\substack{g,\\ A\text{'s random bits}}}{\mathbf{Pr}} [A(g) = YES] \right| \leq \frac{1}{n^{\omega(1)}}$$

where g is a function picked at random from all functions $\{h | h : \{-1,1\}^n \to \{-1,1\}\}$.

Corollary 2.9 *Psuedo-random function generators $\notin AC^0$.*

Proof: Suppose that $f \in \text{AC}^0$. Then we can construct a P.P.T adversary A such that A can tell f apart from a truly random function with non-negligible probability. Consider an adversary A that picks at random $\mathbf{X} \in \{-1,1\}^n$, and $i \in [n]$. A is given an oracle to g, it queries g at $\mathbf{X}$ and $\mathbf{X}^{(i)}$. A outputs YES iff $g(\mathbf{X}) \neq g(\mathbf{X}^{(i)})$.

If g is truly random, then $\mathbf{Pr}[\text{A outputs YES}] = 1/2$; if g is in AC^0, then $\mathbf{Pr}[\text{A outputs YES}] \leq \mathbb{I}(g)/n = \text{poly}\log(n)/n$. □

After seeing all these applications of the LMN theorem, we now show how to prove it. To prove the LMN theorem, we need the the following tools:

Observation 2.10 *A depth-w Decision Tree (DT) is expressible as a width-w DNF or as a width-w CNF.*

Lemma 2.11 *Let $f : \{-1,1\}^n \to \{-1,1\}$, let $(\mathbf{I}, \mathbf{X})$ be a random restriction with $*$-probability ρ. Then $\forall d \geq 5$,*

$$\sum_{|U|\geq 2d/\rho} \widehat{f}(U)^2 \leq 2 \underset{(\mathbf{I},\mathbf{X})}{\mathbf{Pr}} [\textit{DT-depth}(f_{\mathbf{X}\to\overline{\mathbf{I}}}) > d]$$

Now using Lemma 2.11, we can prove the LMN theorem.

Proof:(LMN.)

Claim 2.12

$$\underset{\mathbf{I},\mathbf{X}}{\mathbf{Pr}} [\textit{DT-depth}(f_{\mathbf{X}\to\overline{\mathbf{I}}}) > w] \leq s \cdot 2^{-w}$$

The above claim in combination with Lemma 2.11 would complete the proof. We now show why the claim is true.

Observe that we can view choosing random restriction with $*$-probability $(\frac{1}{10w})^{D-1}$ as the following:

- First choose a random restriction with $*$-probability $\frac{1}{10w}$.
- Further choose a random restriction ont he surviving variables with probability $\frac{1}{10w}$.
- ...

Repeat the above $D-1$ times.

After the first restriction, due to Håstad's Switching Lemma, for any level 2 circuit, the probability that it doesn't turn into a depth-w DT can be bounded as below:

$$\mathbf{Pr}[\text{Doesn't turn into a depth } w \text{ DT}] \leq \left(\frac{5w}{10w}\right)^w = 2^{-w}$$

Due to Observation 2.10, we can express a depth-w DT as a width-w DNF or CNF. Using this, we can transform the bottom two layers of circuit using the opposite of what they were before, i.e., CNF to DNF, and DNF to CNF. This will succeed except with probability $2^{-w}\times$ (number of level 2 gates).

Now the 2nd lowest layer and the 3rd lowest layer will have the same type of gates, so we can collapse them into a single layer. Observe that this operation preserves the bottom fan-in, because the resulting CNF or DNF has width w. So we repeat this operation $D-1$ times, and the probability that the resulting circuit is not a depth-w DT can be bounded as below:

$$\mathbf{Pr}[\text{DT-depth of resulting circuit} > w] \leq \left(\begin{array}{l}\text{number of level 2 gates}\\ +\text{number of level 3 gates}\\ +\text{number of level 2 gates}\\ +\ldots\end{array}\right) \times 2^{-w} \leq s \cdot 2^{-w}$$

□

3 Learning Juntas

We now study how to learn juntas. Let $\mathcal{C}_r = \{f : \{-1,1\}^n \to \{-1,1\}, \text{and } f \text{ is an } r\text{-junta}\}$ denote the family of r-juntas. Note that $\mathcal{C}_{\log n} \subseteq \{\text{poly-size DTs}\} \subseteq \{\text{poly-size DNFs}\}$.

Remark 3.1 *To learn $\mathcal{C}_r$, it suffices for the algorithm to identify the r relevant variables. Since then we can just draw $O(r2^r)$ random examples and with high probability, learn the entire truth table (of size 2^r).*

Observation 3.2 *If f is an r-junta, then every Fourier coefficient of f is either 0 or $\geq 2^{-r}$ in absolute value. This is straightforward directly from the definition of Fourier coefficients, and the fact that f only depend on r variables.*

Fact 3.3 *If $\widehat{f}(S) \neq 0$, then all variables in S are relevant.*

Lecture 9: Learning DNF, AC0, Juntas

Proof: Suppose $\widehat{f}(S) \neq 0$, but there exists $i \in S$ irrelevant, according to the definition, $\widehat{f}(S) = \mathbf{E}_{\mathbf{x}}[f(\mathbf{x})\mathbf{X}_S]$. But for any $\mathbf{x}$, consider $\mathbf{x}^{(i)}$, we have:

$$f(\mathbf{x}) \cdot \mathbf{X}_S = -f(\mathbf{x}^{(i)}) \cdot \mathbf{X}_S^{(i)}$$

Therefore, everything cancels out in $\widehat{f}(S)$, and $\widehat{f}(S) = 0$. □

Using the above facts, we give one idea for learning juntas, and show that r-juntas are learnable in time $\mathrm{poly}(n, 2^r) \cdot n^r$.

- Estimate $\widehat{f}(\emptyset)$.
- Estimate $\widehat{f}(S)$ for all $|S| = 1$ up to accuracy $\frac{2^{-r}}{4}$. If we find an S such that $\widehat{f}(S) \neq 0$, then we know that all variables in S are relevant. Note that this takes time $\mathrm{poly}(n, 2^r)\binom{n}{1}$.
- Estimate $\widehat{f}(S)$ for all $|S| = 2$ up to accuracy $\frac{2^{-r}}{4}$. If we find an S such that $\widehat{f}(S) \neq 0$, then we know that all variables in S are relevant. Note that this takes time $\mathrm{poly}(n, 2^r)\binom{n}{2}$.
- Do the above for S of size $3, 4, \ldots, r$.

Observation 3.4 *The above gives us a* $\mathrm{poly}(n, 2^r) \cdot n^r$*-time algorithm for learning* r*-juntas.*

In the next lecture, we shall improve the above result. In particular, we will ask the question, what kind of functions f can have $\widehat{f}(S) = 0$ for all $1 \leq |S| \leq d$? In particular, if f is not such a function, then by step d in the above algorithm, we will have found a relevant variable. If not, i.e., $\widehat{f}(S) = 0$ for all $1 \leq |S| \leq d$, we will show an algorithm for learning such functions.