17.1 Learning Constant Depth Circuits with MAJ gates

Learning Constant Depth Circuits [LMM '89] says that we can learn size s depth d circuits with n variables to accuracy ϵ in $n^{\log(o(d)\frac{s}{\epsilon})}$ time via a low-degree algorithm. In this lecture, we investigate the question "what is the most interesting concept class we can learn (in a reasonable amount of time)." Clearly we cannot learn depth $\log n$ circuits with the LMN algorithm.

17.2 Circuits with Majority Gates

Here is an interesting concept class that we will show is learnable. Take a circuit in AC^0 (the class of polynomial-sized constant depth circuits) and stick a single majority ("MAJ") gate at the top. A MAJ gate returns true when the majority of its inputs are true. Before discussing how we can learn this concept class, let's provide some reasons for why this is an interesting concept class to investigate.

1. The gate MAJ is not in AC^0.

2. Combines two different concept classes. AC^0 is a circuit, while MAJ is a halfspace. In fact, the techniques we use will work for <u>any</u> threshold function with polynomially bounded weights.

3. MAJ gates are powerful. If we allow polynomially many, then we can compute many interesting functions

17.3 Learning Circuits with Majority Gates

MAJ is a halfspace, so we can use Fourier learning techniques to learn MAJ. The following captures most of the Fourier spectrum of MAJ:

$$\sum_{|s|\leq\frac{1}{\epsilon^2}} \hat{f}(s)^2 = \Theta(1-\epsilon)$$

Learning MAJ therefore requires time $n^{o(\frac{1}{\epsilon^2})}$ via the low-degree algorithm.

Let f, g be $\{+1, -1\}$ functions, and call $E_D[f \cdot g]$ the correlation of f and g. If $E_D[f \cdot g] = \gamma$, this implies that

$$\Pr_{x\in D}[f(x)=g(x)] \geq \frac{1}{2} + \frac{\gamma}{2}.$$

Claim 1 *Suppose that for $a_i \in \mathbb{Z}$,*

$$g = SIGN(\sum_{i=1}^{s} a_i h_i),$$

and that $E_D[g] \leq 1/s$. Then $\forall D$ on $\{-1, +1\}^n, \exists i$ s.t.

$$E_D[g \cdot h_i] \geq \frac{1}{\sum_{i=1}^{s} |a_i|}.$$

For instance, if $g = \text{MAJ}(h_1, ..., h_s)$ this implies $\exists h_i$ s.t. $E_D[g \cdot h_i] \geq \frac{1}{s}$. Intuitively, this theorem tells us that g is non-trivially correlated with at least on of the h_i.

Proof:

$$
\begin{aligned}
1 &= E_D[f \cdot f] \\
&\leq E_D[f \cdot \left(\sum_{i=1}^{s} a_i h_i\right)] \\
&\leq \sum_{i=1}^{s} |a_i| |E_D[f \cdot h_i]| \\
&\leq \max_i E_D[f \cdot h_i] \cdot \sum_{i=1}^{s} |a_i|.
\end{aligned}
$$

∎

Fix D on $\{-1,+1\}$.
Let $f = \text{MAJ}(C_1, ..., C_S)$. It would be nice if we could learn f by first learning each C_i and then learning MAJ. LMN tells us that we can learn each C_i, because for any depth d size s circuit C,

$$\sum_{|A| \geq k} \hat{C}(A)^2 \leq 2S \cdot 2^{-k^{\frac{1}{d}}/20}$$

Unfortunately, we only see the output of MAJ, so we cannot apply LMN directly to learn the circuits C_i. However claim 1 tells us that at least one of these circuits is correlated with MAJ.

Let $k = O(\log(s^3 2^n \cdot L_\infty(D))^d)$,
Let C_D be the circuit with $\frac{1}{S}$ correlation with f.
Now let $g = \sum_{|A| \leq k} \hat{C}_d(A) \chi_A$. Then,

$$
\begin{aligned}
E_U[(g - C_D)^2] &\leq \frac{1}{4s^2 \cdot 2^n L_\infty(D)} \\
E_D[(g - C_D)^2] &\leq \frac{1}{4s^2} \\
E_D[|g - C_D|] &\leq \frac{1}{2s}.
\end{aligned}
$$

$$
\begin{aligned}
E_D[f \cdot g] &= E_D[f \cdot C_D] + E_D[f(g - C_D)] \\
&\geq \frac{1}{S} - E_D[|g - C_D|] \\
&\geq \frac{1}{S} - \frac{1}{2S} \\
&= \frac{1}{2S}.
\end{aligned}
$$

So g has non-negligible correlation with the output of MAJ. This is going to imply that there exists some χ_A such that

$$
E_D[f \cdot \chi_A] \geq \frac{1}{2Sn^k}.
$$

Proof:

$$
\begin{aligned}
\sum_{|A| \leq k} |\hat{C}_D(A)| \cdot |E_D[f \cdot \chi_A]| &\geq \sum_{|A| \leq k} \hat{C}_D(A) \cdot E_D[f \cdot \chi_A] \\
&= E_D[f \cdot \sum_{|A| \leq k} \hat{C}_D(A)\chi_A] \\
&= E_D[f \cdot g] \\
&\geq \frac{1}{2S}.
\end{aligned}
$$

implies there exists A such that $|E_D[f \cdot \chi_A]| \geq \frac{1}{2Sn^k}$. ∎

The weak learning algorithm runs in time $n^{O(k)}$ and outputs χ_A with correlation $\frac{1}{2Sn^k}$. We apply a Boosting Algorithm:

- for each D_i, $L_\infty(D_i) = O(\frac{1}{\epsilon})$.

- Number of iterations is poly($\frac{1}{\gamma}, \frac{1}{\epsilon}, n$,time of WL algorithm)

- Output is MAJ of the WL hypothesis.

The bottom line is that we can learn a constant depth circuit with a single MAJ gate in time polynomial in $n^{\log(\frac{S^3}{\epsilon})^d}$

17.4   A Theorem about Majority Gates

This theorem is due to [Beigel], and says that we can represent a circuit with m MAJ gates by a circuit with a single MAJ gate, but at a cost to size and depth. Let f be a circuit c of size s and depth d with m MAJ gates. Then there exists a circuit of size $2^{m \cdot O(\log s)^{2d+1}}$ and depth $d + 2$ such that $c = f$, and c has 1 MAJ gate at the root.

This is the most expressive circuit class that we know how to learn in quasipolynomial time with respect to the uniform distribution.