Learnability of AC^0 (lecture notes)

Course: Circuit Complexity, Autumn 2024, University of Chicago
Instructor: William Hoza (`williamhoza@uchicago.edu`)

In these notes, it will be convenient to encode bits using the values $\{\pm 1\}$ instead of $\{0,1\}$.

Definition 1 (Closeness under the uniform distribution)**.** Let $C, C' \colon \{\pm 1\}^n \to \{\pm 1\}$. We define

$$\mathrm{dist}(C, C') = \Pr_{x \in \{\pm 1\}^n}[C(x) \neq C'(x)].$$

Our goal in these notes is to prove the following theorem:

Theorem 1 (Learnability of AC^0)**.** *Let $C \colon \{\pm 1\}^n \to \{\pm 1\}$ be an "unknown" AC^0_d circuit of size at most S. Suppose we are given access to an unlimited supply of independent samples of the form $(x, C(x))$ where $x \in \{\pm 1\}^n$ is drawn uniformly at random. Suppose also that we are given the parameters S and d, as well as $\varepsilon, \delta \in (0,1)$. Then with probability $1-\delta$, we can construct a circuit h such that $\mathrm{dist}(h, C) \leq \varepsilon$ in time $n^{O(\log S)^{d-1} \cdot \log(1/\varepsilon)} \cdot \mathrm{polylog}(1/\delta)$.*

For example, if $d = O(1)$, $S = \mathrm{poly}(n)$, $\varepsilon = 1/\mathrm{poly}(n)$, and $\delta = 2^{-n}$, then the time complexity is $\mathrm{quasipoly}(n)$.

1 Fourier analysis of AC^0

The proof of Theorem 1 is based on *Fourier analysis* of AC^0 circuits. (If you're not familiar with the Fourier analysis of Boolean functions, then you should read sections 1.1-1.4 of O'Donnell's book [O'D14] before reading the rest of these lecture notes.) Indeed, the Fourier analysis involved in the proof of Theorem 1 is arguably more important than Theorem 1 itself. The main ingredient in the proof of Theorem 1 is the following "Fourier tail bound" for AC^0.

Theorem 2 (Fourier tail bound for AC^0)**.** *Let $C \colon \{\pm 1\}^n \to \{\pm 1\}$ be an AC^0_d circuit of size S. Then for every $k \in \mathbb{N}$, we have*

$$\sum_{\substack{S \subseteq [n] \\ |S| \geq k}} \widehat{C}(S)^2 \leq 2 \cdot 2^{-k/O(\log S)^{d-1}}.$$

The first Fourier tail bound for AC^0 was proven by Linial, Mansour, and Nisan [LMN93], and it is sometimes called the "LMN theorem." Theorem 2 is a quantitative improvement due to Tal [Tal17]. We will prove Theorem 2 in the upcoming sections. Before we do so, let us show how to use Theorem 2 to prove Theorem 1.

Proof of Theorem 1 given Theorem 2. Let Let $k, t \in \mathbb{N}$ be parameters that we will choose later. Recall that for each $S \subseteq [n]$, the Fourier coefficient $\widehat{C}(S)$ is given by $\widehat{C}(S) = \mathbb{E}_x[C(x) \cdot \chi_S(x)]$. For each $S \subseteq [n]$ with $|S| < k$, we use t random labeled examples $(x^{(1)}, C(x^{(1)})), \dots, (x^{(t)}, C(x^{(t)}))$ to construct an estimate $\widehat{\phi}(S)$ for $\widehat{C}(S)$ as follows:

$$\widehat{\phi}(S) = \frac{1}{t} \sum_{i=1}^{t} C(x^{(i)}) \cdot \chi_S(x^{(i)}).$$

Now define

$$\phi(x) = \sum_{\substack{S \subseteq [n] \\ |S| < k}} \widehat{\phi}(S) \cdot \chi_S(x), \qquad h(x) = \mathrm{sign}(\phi(x)).$$

To prove that this works, observe that

$$\begin{aligned}\mathrm{dist}(C,h) = \Pr_{x\in\{\pm1\}^n}[C(x)\neq h(x)] &\le \Pr_{x\in\{\pm1\}^n}[|C(x)-\phi(x)|\ge 1]\\ &\le \mathop{\mathbb{E}}_{x\in\{\pm1\}^n}[(C(x)-\phi(x))^2]\\ &= \sum_{\substack{S\subseteq[n]\\|S|<k}}(\widehat{\phi}(S)-\widehat{C}(S))^2+\sum_{\substack{S\subseteq[n]\\|S|\ge k}}\widehat{C}(S)^2 \qquad \text{(Parseval.)}\end{aligned}$$

By Theorem 2, the second term is at most $\varepsilon/2$, provided that we choose a suitable value $k = O(\log S)^{d-1}\cdot\log(1/\varepsilon)$. Regarding the first term, for each fixed S, we can apply Hoeffding's inequality to get

$$\Pr\left[\left|\widehat{\phi}(S)-\widehat{C}(S)\right| > \sqrt{\frac{\varepsilon}{2n^k}}\right] \le 2\exp(-\Omega(\varepsilon t/n^k)).$$

If we choose a suitable value $t = O(n^k\cdot\log(n^k/\delta)/\varepsilon)$, then this failure probability is less than δ/n^k. By the union bound, we may assume that $|\widehat{\phi}(S)-\widehat{C}(S)|\le\sqrt{\frac{\varepsilon}{2n^k}}$ for all $S\subseteq[n]$ with $|S|<k$, and hence

$$\sum_{\substack{S\subseteq[n]\\|S|<k}}(\widehat{\phi}(S)-\widehat{C}(S))^2\le\frac{\varepsilon}{2}.$$ □

2 AC^0 circuits become low-degree functions under restrictions

In a previous class, we discussed the AC^0 Criticality Theorem, which describes the effect of random restrictions on AC^0 circuits.

Theorem 3 (AC^0 Criticality Theorem)**.** *Let C be a size-S AC^0_d circuit, let $p\in(0,1)$, and let $D\in\mathbb{N}$. Then*

$$\Pr_{\rho\sim R_p}[\mathsf{DTDepth}(C|_\rho)\ge D]\le(p\cdot O(\log S)^{d-1})^D.$$

Theorem 3 is the *only* fact about AC^0 circuits that we will use to prove Theorem 2. All of the other steps of the proof are generic and apply to arbitrary Boolean functions.

The reason Theorem 3 is helpful for us is that low-depth decision trees have no high-degree Fourier mass, as we will prove momentarily. On the other hand, we will prove in the next section that random restrictions do not have a huge effect on a function's Fourier tails. This will enable us to conclude that the circuit must have had bounded Fourier tails to begin with, even before applying the random restriction.

Proposition 1 (Shallow decision trees have low degree)**.** *Let $T\colon\{\pm1\}^n\to\{\pm1\}$ be a decision tree of depth D. Then $\deg(T)\le D$, where $\deg(T)$ denotes the Fourier degree of T, i.e., the degree of T as a multilinear real polynomial.*

Proof. We can write T in the form $T(x)=\sum_{\ell\in L}c_\ell\cdot T_\ell(x)$, where L is the set of leaves, c_ℓ is the output value at leaf ℓ, and $T_\ell(x)$ indicates whether the tree reaches leaf ℓ on input x. Each function T_ℓ depends on at most D variables, hence $\deg(T_\ell)\le D$, hence $\deg(T)\le D$. □

3 Random restrictions have little effect on Fourier tails

To complete the proof of Theorem 2, we need to bound the effect of random restrictions on the Fourier weights. By Parseval's theorem, we have $\sum_{S\subseteq[n]}\widehat{C}(S)^2=1$. Consequently, we can interpret $\widehat{C}(S)^2$ as a probability. We define the *spectral sample* $\mathcal{S}_C$ to be the probability distribution over subsets of $[n]$ in which the probability of getting any particular set S is $\widehat{C}(S)^2$. Thus, the Fourier tail bound we are trying to prove (Theorem 2) can be rephrased as follows:

$$\Pr_{S\sim\mathcal{S}_C}[|S|\geq k]\leq 2\cdot 2^{-k/O(\log S)^{d-1}}.$$

The key to proving it is the following lemma, which says that the operation of drawing a spectral sample "commutes with" the operation of applying a random restriction.

Lemma 1 (Spectral sample after a random restriction)**.** *Let $C\colon\{\pm 1\}^n\to\{\pm 1\}$. The following two distributions over subsets of $[n]$ are identical.*

1. *Sample $\rho\sim R_p$, then sample $S\sim\mathcal{S}_{C|_\rho}$, then output S.*
2. *Sample $T\sim\mathcal{S}_C$, then sample $\rho\sim R_p$, then output $T\cap\rho^{-1}(\star)$.*

Proof. If ρ is a restriction and x is a completion of ρ, then we have

$$C(x)=\sum_{T\subseteq[n]}\widehat{C}(T)\cdot\chi_T(x)=\sum_{T\subseteq[n]}\widehat{C}(T)\cdot\chi_{T\cap\rho^{-1}(\{0,1\})}(x)\cdot\chi_{T\cap\rho^{-1}(\star)}(x).$$

Consequently, for any $S\subseteq[n]$, the Fourier coefficient $\widehat{C|_\rho}(S)$ is given by the following formula.

$$\widehat{C|_\rho}(S)=\sum_{U\subseteq[n]}\widehat{C}(S\cup U)\cdot\chi_U(x)\cdot 1[S\subseteq\rho^{-1}(\star)\text{ and }U\subseteq\rho^{-1}(\{0,1\})].$$

Squaring the equation above, we get

$$\widehat{C|_\rho}(S)^2=\sum_{U,U'\subseteq[n]}\widehat{C}(S\cup U)\cdot\widehat{C}(S\cup U')\cdot\chi_{U\Delta U'}(x)\cdot 1[S\subseteq\rho^{-1}(\star)\text{ and }U,U'\subseteq\rho^{-1}(\{0,1\})],$$

where $U\Delta U'$ is the symmetric difference between U and U'. All of the above holds for any fixed restriction ρ and any completion x of ρ. If ρ is a random restriction sampled from R_p and x is a uniform random completion of ρ, then in expectation, we have

$$\mathbb{E}\left[\widehat{C|_\rho}(S)^2\right]=\sum_{U,U'\subseteq[n]}\widehat{C}(S\cup U)\cdot\widehat{C}(S\cup U')\cdot\mathbb{E}\left[\chi_{U\Delta U'}(x)\cdot 1[S\subseteq\rho^{-1}(\star)\text{ and }U,U'\subseteq\rho^{-1}(\{0,1\})]\right].$$

The completion x and the star-set $\rho^{-1}(\star)$ are independent, so we can exchange the expectation with the product:

$$\mathbb{E}\left[\widehat{C|_\rho}(S)^2\right]=\sum_{U,U'\subseteq[n]}\widehat{C}(S\cup U)\cdot\widehat{C}(S\cup U')\cdot\mathbb{E}[\chi_{U\Delta U'}(x)]\cdot\Pr[S\subseteq\rho^{-1}(\star)\text{ and }U,U'\subseteq\rho^{-1}(\{0,1\})].$$

Nontrivial character functions have expectation zero, so the equation above simplifies to

$$\begin{aligned}\mathbb{E}\left[\widehat{C|_\rho}(S)^2\right]&=\sum_{U\subseteq[n]}\widehat{C}(S\cup U)^2\cdot\Pr[S\subseteq\rho^{-1}(\star)\text{ and }U\subseteq\rho^{-1}(\{0,1\})]\\&=\sum_{T\subseteq[n]}\widehat{C}(T)^2\cdot\Pr[S=T\cap\rho^{-1}(\star)].\end{aligned}$$

The left-hand side in the equation above is the probability of getting S under distribution 1 in the lemma statement. The right-hand side is the probability of getting S under distribution 2 in the lemma statement. □

Proof of Theorem 2. On the one hand, by Proposition 1 and Theorem 3, there is a value $p = 1/O(\log S)^{d-1}$ such that for every $D \in \mathbb{N}$, we have

$$\Pr_{\substack{\rho\sim R_p \\ S\sim\mathcal{S}_{C|_\rho}}}[|S| \geq D] \leq \Pr_{\rho\sim R_p}[\mathsf{DTDepth}(C|_\rho) \geq D] \leq 2^{-D}.$$

On the other hand, by Lemma 1, we have

$$\Pr_{\substack{\rho\sim R_p \\ S\sim\mathcal{S}_{C|_\rho}}}[|S| \geq D] = \mathop{\mathbb{E}}_{T\sim\mathcal{S}_C}\left[\Pr_{\rho\sim R_p}[|T \cap \rho^{-1}(\star)| \geq D]\right].$$

For any fixed set $T \subseteq [n]$, we expect $|T \cap \rho^{-1}(\star)| \approx p \cdot |T|$. Indeed, one can show that

$$\Pr\left[|T \cap \rho^{-1}(\star)| \geq \lfloor\, p \cdot |T|\, \rfloor\right] \geq 1/2.$$

(Note that such a statement amounts to bounding the median of the binomial distribution.[1]) Therefore,

$$\mathop{\mathbb{E}}_{T\sim\mathcal{S}_C}\left[\Pr_{\rho\sim R_p}\left[|T \cap \rho^{-1}(\star)| \geq \lfloor pk \rfloor\right]\right] \geq \Pr_{T\sim\mathcal{S}_C}[|T| \geq k] \cdot \frac{1}{2}.$$

Rearranging, we get $\Pr_{T\sim\mathcal{S}_C}[|T| \geq k] \leq 2 \cdot 2^{-\lfloor pk \rfloor}$. If $pk \geq 2$, then this is at most $2 \cdot 2^{-pk/2}$, and if $pk \leq 2$, then trivially $\Pr_{T\sim\mathcal{S}_C}[|T| \geq k] \leq 2 \cdot 2^{-pk/2}$. □

References

[LMN93] Nathan Linial, Yishay Mansour, and Noam Nisan. "Constant depth circuits, Fourier transform, and learnability". In: *J. Assoc. Comput. Mach.* 40.3 (1993), pp. 607–620. ISSN: 0004-5411. DOI: 10.1145/174130.174138. URL: https://doi.org/10.1145/174130.174138.

[O'D14] Ryan O'Donnell. *Analysis of Boolean functions*. Available online at https://arxiv.org/abs/2105.10386. Cambridge University Press, New York, 2014, pp. xx+423. ISBN: 978-1-107-03832-5. DOI: 10.1017/CBO9781139814782.

[Tal17] Avishay Tal. "Tight Bounds on the Fourier Spectrum of AC0". In: *Proceedings of the 32nd Computational Complexity Conference (CCC)*. Ed. by Ryan O'Donnell. Vol. 79. 2017, 15:1–15:31. DOI: 10.4230/LIPIcs.CCC.2017.15.

[1] An alternative and more elementary approach is to use Cantelli's inequality to prove $\Pr[|T \cap \rho^{-1}(\star)| \geq \lfloor pk/2 \rfloor] \geq 1/3$.