CS 6817: Special Topics in Complexity Theory **Spring 2025**

Lecture 9: Feb 20, 2025

Lecturer: Eshan Chattopadhyay *Scribe: Noam Ringach*

1 Decision Trees

Recall our notation that decision trees have size DT_{size} and depth DT_{depth}, and that these are complexity measures we can associate with a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ by considering the smallest decision tree computing f. Using these complexity measures, we will bound the Fourier tail of f.

Definition 1.1. *Recall the Fourier level and Fourier tail definitions.*

$$W^k[f] = \sum_{\substack{S\subseteq[n]\\|S|=k}} \hat{f}(S)^2$$
$$W^{\geq k}[f] = \sum_{\substack{S\subseteq[n]\\|S|\geq k}} \hat{f}(S)^2$$

Claim 1.2. *If $DT_{depth}(f) = d$, then $W^{\geq d+1}[f] = 0$. In other words,* $\deg(f) \leq d$.

Proof. Let T be the decision tree computing f with $\text{depth}(T) \leq d$. We're going to write all paths down the decision tree as a monomial. That is, since every node in a decision tree is labeled by some x_i, if the path taken from the root to some leaf is $(x_{i_1}, x_{i_2}, \ldots, x_{i_d})$, we would consdier the monomial $x_{i_1}x_{i_2}\cdots x_{i_d}$. Let $\mathcal{P} = \{\text{all root to leaf paths in } T\}$. So we can write

$$f(x) = \sum_{P\in\mathcal{P}} \ell_P \cdot 1_P(x)$$

where ℓ_P is the label of the leaf at the end of P, and $1_P(x)$ is 1 if and only if P is the path taken when x is given as an input to f and 0 otherwise. This decomposition works because only one indicator will ever be 1 while the rest are 0 since all paths are unique (we're in a tree), and we're giving the correct value to that path.

Now that we've decomposed f in this way, we can focus on the Fourier decomposition of $1_P(x)$. Let $b_{1,P}, \ldots, b_{d,P}$ be the edge labels given by P (we assume wlog that P uses d edges). The condition for $1_P(x)$ then is $1_P(x) = (x_{i_{1,P}} = b_{1,P}) \wedge (x_{i_{2,P}} = b_{2,P}) \wedge \cdots \wedge (x_{i_{d,P}} = b_{d,P}) = \prod_{j=1}^{d} \left(\frac{1+x_{i_{j,P}}b_{j,P}}{2}\right)$ where $i_{1,P}, \ldots, i_{d,P}$ are the indices of the d variables used along the path of P. Therefore,

$$f(x) = \sum_{P\in\mathcal{P}} \ell_P \cdot \prod_{j=1}^{d} \left(\frac{1+x_{i_{j,P}}b_{j,P}}{2}\right).$$

From this we see that there is no monomial with degree more than d, so $\deg(f) \leq d$, as claimed. □

Next we claim a similar result but for the size of a decision tree.

Claim 1.3. *If $DT_{size}(f) = s$, then $W^{\geq k+1}[f] \leq 4\varepsilon$ where $k = \log(s/\varepsilon)$.*

Proof. Well, we only really know how to do one thing, which is to argue about the depth of a decision tree, so we can try to do a similar argument here where we limit the depth of the decision tree by ignoring long paths. Suppose T computes f with $\text{size}(T) = s$. Define T' to be the same as T except with any path of depth $> k$ truncated with a leaf label of -1. Therefore, by our previous claim we know that $W^{\geq k+1}[T'] = 0$ (here we are overloading notation by letting T' itself be the function represented by the tree T').

This is close to what we want, but we unfortunately have that $T \neq T'$ unless T already has depth at most k. How far is T from T'? We can compute

$$\Pr_{x \sim \{-1,1\}^n}[T'(x) \neq T(x)] \leq \frac{s}{2^k}$$

since there are at most s leaves that we cut off by truncating and getting to such a path requires taking a path of length k which has probability $\frac{1}{2^k}$, so we just union bound over all such paths. Hence, by assumption that $k = \log(s/\varepsilon)$, we get $\Pr_{x \sim \{-1,1\}^n}[T'(x) \neq T(x)] \leq \varepsilon$.

Now let's consider the error function $\mathcal{E} : \{-1,1\}^n \to \mathbb{R}$ where $\mathcal{E}(x) = T'(x) - T(x)$.

Observation 1.4. *Observe that $W^{\geq k+1}[\mathcal{E}] = W^{\geq k+1}[T]$ since $\deg(T') \leq k$ so the degree $\geq k+1$ coefficients of $\mathcal{E}$ only come from T. We'll formally prove this below.*

Proof of observation. By construction, for any $S \subseteq [n]$ with $|S| \geq k+1$, we have $\widehat{\mathcal{E}}(S) = -\widehat{f}(S)$. And by Parseval we have $\|\mathcal{E}\|_2^2 = \sum_{S \subseteq [n]} \widehat{\mathcal{E}}(S)^2$. Therefore

$$\mathop{\mathbb{E}}_{x \sim \{-1,1\}^n}[(T(s) - T'(x))^2] \leq 4\varepsilon,$$

which bounds all Fourier coefficients, implying that $W^{\geq k+1}[\mathcal{E}] \leq 4\varepsilon$. □

Noticing that $T(x) = T'(x) - \mathcal{E}(x)$ completes the proof.

□

2 PAC (Probably Approximately Correct) Learning Application

We now focus on PAC learning and whether it's possible to estimate Fourier coefficients efficiently. First, we formally define PAC learning:

Definition 2.1 (PAC Learning)**.** *We're given a* concept class*, a family of boolean functions, $\mathcal{F}$ and one of two possible modesl for getting information about a function $f \in \mathcal{F}$. In the* random example model*, we get $\{(x^{(i)}, f(x^{(i)})\}_{i=1,2,\ldots}$ where we're guaranteed that each $x^{(i)} \sim \{-1,1\}^n$. Alternatively, we can consider the query model where we get to choose the $x^{(i)}$'s (adaptively), which is clearly easier to learn.*

The goal is to create a randomized algorithm $\mathcal{A}$ that outputs a function $h : \{-1,1\}^n \to \{-1,1\}$ called our hypothesis*. We do not require that $h \in \mathcal{F}$ (if $h \in \mathcal{F}$, then this is called* proper *learning), only that with probability $1 - \delta$ (over the randomness of $\mathcal{A}$) for some small $\delta > 0$, we get*

$$\Pr_{x \sim \{-1,1\}^n}[h(x) = f(x)] \geq 1 - \varepsilon.$$

A natural question to ask is whether functions with small Fourier tails are easy to learn. The following theorem says yes!

Theorem 2.2 (Low degree algorithm)**.** *Suppose $\mathcal{F}$ is such that for any $f \in \mathcal{F}$, $W^{\geq k}[f] \leq \varepsilon$. Then $\mathcal{F}$ is PAC learnable in the random example model in time* $\text{poly}(n^k, 1/\varepsilon, \log(1/\delta))$.

Corollary 2.3. *Size s decision trees are learnable in* $\mathrm{poly}(n^{\log(s/\varepsilon)}, 1/\varepsilon, \log(1/\delta))$ *time.*

The key ingredient to the low degree algorithm is the algorithm FOURIER.

Definition 2.4. FOURIER : $2^{[n]} \to \mathbb{R}$ *is an algorithm given access to random examples* $(x^{(i)}, f(x^{(i)}))$ *for* $x^{(i)} \sim \{-1,1\}^n$ *with the goal of approximating* $\widehat{f}(S)$ *to accuracy* ε *with probability* $1-\delta$.

Algorithm:

1. *Get* t *samples* $\{x^{(i)}, f(x^{(i)})\}_{i=1}^t$
2. *Compute the empirical mean* $\widehat{f}(S) = \frac{1}{t}\left(\sum_{i=1}^t f(x^{(i)}) \cdot \chi_S(x^{(i)})\right)$

To prove our algorithm works, we will need Hoeffding's inequality.[1] We state the specific version we need here.

Lemma 2.5 (Hoeffding's Inequality)**.** *Let* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t$ *be independent random variables taking values in* $[-1,1]$ *such that for all* $i \in [t]$ *we have* $\mu = \mathbb{E}[\boldsymbol{X}_i]$. *Then*

$$\Pr\left[\left|\frac{1}{t}\sum_{i=1}^t \boldsymbol{X}_i - \mu\right| > \varepsilon\right] \leq 2e^{-\Omega(t\varepsilon^2)}$$

We can now prove our desired claim.

Claim 2.6. *The algorithm for FOURIER works as claimed with accuracy* ε *with probability* $1-\delta$.

Proof. Let $y_i = f(x^{(i)})\chi_S(x^{(i)})$ for $i \in [t]$, so $\mathbb{E}_{x^{(i)}\sim\{-1,1\}^n}[y_i] = \widehat{f}(S)$. Then we can use Hoeffding's inequality to compute

$$\Pr\left[\left|\frac{1}{t}\sum_{i=1}^t y_i - \widehat{f}(S)\right| > \varepsilon\right] \leq 2e^{-\Omega(t\varepsilon^2)},$$

meaning that $t = O\left(\frac{1}{\varepsilon^2}\log(1/\delta)\right)$. □

In order to prove Theorem 2.2, all we need to do is union bound over all low degree S.

[1] The proof of Hoeffding's inequality follows that of Chernoff's bound, which can be found in the Lecture notes for CS 4850, and a specific proof of Hoeffding's inequality can be found in its Wikipedia page