Fourier Concentration, Decision Trees, and Learning Theory (lecture notes)

Course: Analysis of Boolean Functions, Autumn 2025, University of Chicago
Instructor: William Hoza (`williamhoza@uchicago.edu`)

1 Learnability of low-degree functions

Definition 1.1 (Learning from random examples)**.** Let $\mathcal{C}$ be a class of functions $f\colon \{\pm 1\}^n \to \{\pm 1\}$. We say that $\mathcal{C}$ *can be learned from random examples* with error ε in time T if there is a time-T randomized algorithm with the following behavior.

- Initially, the algorithm is given some parameters that are part of the definition of $\mathcal{C}$. These parameters will be clear from context.
- At any time, the algorithm can press a button to receive $(x, f(x))$ for a uniform random $x \in \{\pm 1\}^n$.
- With probability at least $9/10$, the algorithm outputs a hypothesis $h\colon \{\pm 1\}^n \to \{\pm 1\}$ (represented by a Boolean circuit) such that $\mathrm{dist}(f, h) \leq \varepsilon$.

Admittedly, this learning model has some major weaknesses. We assume that the input distribution is uniform; we assume that the examples are noiseless; the output hypothesis h is not guaranteed to belong to the concept class $\mathcal{C}$. Still, it's an interesting and appealing model.

Let us prove that depth-k decision trees can be learned from random examples with error 0 in time $n^{O(k)}$. More generally, we will show how to handle degree-k Boolean functions.

Lemma 1.2. *If f is a depth-k decision tree, then* $\deg(f) \leq k$.

Proof. Express f as a sum over leaves. The contribution from a single leaf is a function of at most k variables, hence it has degree at most k. □

Lemma 1.3. *For any $f\colon \{\pm 1\}^n \to [-1, 1]$, given any $S \subseteq [n]$, using random examples, we can estimate $\widehat{f}(S)$ to within $\pm\varepsilon$ except with probability δ in time* $\mathrm{poly}(n, 1/\varepsilon, \log(1/\delta))$.

Proof. This follows from Hoeffding's inequality. □

Theorem 1.4. *The class of Boolean functions of degree at most k can be learned from random examples with error 0 in time $n^{O(k)}$.*

Proof. For each set $S \subseteq [n]$ of size at most k, estimate $\widehat{f}(S)$ to within $\pm 0.1 \cdot 2^{-k}$ with failure probability $n^{-k}/10$. By the union bound, we can assume that all of these estimates succeed.

In Exercise 1, you prove that the Fourier coefficients of a degree-k Boolean function are always integer multiples of 2^{-k}. Therefore, round all the estimated Fourier coefficients to the nearest integer multiple of 2^{-k}. This is the exact Fourier expansion of f. □

The theorem suggests that we should think of $\deg(f)$ as a measure of the "complexity" of f. It turns out that if $k \in \mathbb{N}$ and $\varepsilon \in (0, 1)$ are both held constant, then degree-k Boolean functions can be learned from $O(\log n)$ random examples with error ε [EI22].

2 Learnability of functions that are concentrated at low degree

In this section, we will show that size-s decision trees can be learned from random examples with error ε in time $n^{O(\log(s/\varepsilon))}$.

Lemma 2.1. *If f is a size-s decision tree, then f is ε-close to a decision tree of depth $\log(s/\varepsilon)$.*

Proof. Define

$$g(x) = \begin{cases} f(x) & \text{if } f \text{ makes at most } \log(s/\varepsilon) \text{ queries on } x \\ +1 & \text{otherwise.} \end{cases}$$

For any leaf u at depth at least k in f, the probability of reaching u is at most 2^{-k}, so by the union bound, the probability of reaching a deleted leaf is at most $s \cdot 2^{-\log(s/\varepsilon)} = \varepsilon$. □

Definition 2.2 (Total variation distance)**.** Let D_1, D_2 be probability distributions over the finite set Ω. The *total variation distance* between D_1 and D_2 is defined to be

$$\max_{\mathcal{F} \subseteq \Omega} \left| \Pr_{x \sim D_1}[x \in \mathcal{F}] - \Pr_{x \sim D_2}[x \in \mathcal{F}] \right|.$$

Lemma 2.3. *Let $f, g \colon \{\pm 1\}^n \to \{\pm 1\}$. The total variation distance between the spectral samples $\mathcal{S}_f$ and $\mathcal{S}_g$ is $O(\sqrt{\operatorname{dist}(f,g)})$.*

Proof. Let $\mathcal{F}$ be any collection of subsets of $[n]$. Then

$$\begin{aligned} \left| \Pr_{S \sim \mathcal{S}_f}[S \in \mathcal{F}] - \Pr_{S \sim \mathcal{S}_g}[S \in \mathcal{F}] \right| &= \left| \sum_{S \in \mathcal{F}} \left(\widehat{f}(S)^2 - \widehat{g}(S)^2 \right) \right| \\ &= \left| \sum_{S \in \mathcal{F}} (\widehat{f}(S) - \widehat{g}(S)) \cdot (\widehat{f}(S) + \widehat{g}(S)) \right| \\ &\leq \sum_{S \subseteq [n]} |\widehat{f}(S) - \widehat{g}(S)| \cdot |\widehat{f}(S) + \widehat{g}(S)| \\ &\leq \sqrt{\left(\sum_{S \subseteq [n]} (\widehat{f}(S) - \widehat{g}(S))^2 \right) \cdot \left(\sum_{S \subseteq [n]} (\widehat{f}(S) + \widehat{g}(S))^2 \right)} \qquad \text{(Cauchy-Schwarz)} \\ &= \sqrt{\mathop{\mathbb{E}}_x[(f(x) - g(x))^2] \cdot \mathop{\mathbb{E}}_x[(f(x) + g(x))^2]} \qquad \text{(Parseval)} \\ &\leq \sqrt{4 \operatorname{dist}(f,g) \cdot 4}. \end{aligned}$$

□

Definition 2.4 (Concentration at low degree)**.** Let $f \colon \{\pm 1\}^n \to \mathbb{R}$ and let $0 \leq k \leq n$. We define

$$W^{>k}[f] = \sum_{S \subseteq [n], |S| > k} \widehat{f}(S)^2.$$

We say that (the Fourier spectrum of) f is *ε-concentrated up to degree k* if $W^{>k}[f] \leq \varepsilon$. In the case that f is $\{\pm 1\}$-valued, it is equivalent to say that $\Pr_{S \sim \mathcal{S}_f}[|S| > k] \leq \varepsilon$.

Corollary 2.5. *If $f \colon \{\pm 1\}^n \to \{\pm 1\}$ is a size-s decision tree, then f is ε-concentrated at degree up to $O(\log(s/\varepsilon))$.*

Theorem 2.6 (The Low-Degree Algorithm)**.** *The class of Boolean functions that are ε-concentrated up to degree k can be learned from random examples with error 2ε in time $n^{O(k)} \cdot \operatorname{poly}(1/\varepsilon)$.*

Proof. For each set $S \subseteq [n]$ of size at most k, estimate $\widehat{f}(S)$ to within $\pm\sqrt{\varepsilon/(n+1)^k}$ with failure probability $\frac{1}{10\cdot(n+1)^k}$. By the union bound, we can assume that all of these estimates succeed. Let c_S be the estimate for $\widehat{f}(S)$. Hypothesis:

$$h(x) = \mathrm{sign}\left(\sum_{S\subseteq[n],|S|\leq k} c_S \cdot \chi_S(x)\right).$$

This works, because

$$\mathrm{dist}(f,h) \leq \mathop{\mathbb{E}}_x\left[\left(f(x) - \sum_{S\subseteq[n],|S|\leq k} c_S\cdot\chi_S(x)\right)^2\right] = W^{>k}[f] + \sum_{S\subseteq[n],|S|\leq k}(\widehat{f}(S)-c_S)^2 \leq 2\varepsilon. \qquad \square$$

3 The Kushilevitz-Mansour algorithm

In the previous section, we showed that polynomial-size decision trees can be learned from random examples in quasipolynomial time $n^{O(\log n)}$. It is an open question whether the time complexity can be improved to polynomial. In this section, we will show how to improve the time complexity to polynomial if we are allowed to make *queries* to the unknown function.

Definition 3.1 (Learning from queries)**.** Let $\mathcal{C}$ be a class of functions $f\colon \{\pm1\}^n \to \{\pm1\}$. We say that $\mathcal{C}$ *can be learned from queries* with error ε in time T if there is a time-T randomized algorithm with the following behavior.

- Initially, the algorithm is given some parameters that are part of the definition of $\mathcal{C}$. These parameters will be clear from context.
- At any time, the algorithm can query f at any input $x \in \{\pm1\}^n$ of its choosing.
- With probability at least 9/10, the algorithm outputs a hypothesis $h\colon \{\pm1\}^n \to \{\pm1\}$ such that $\mathrm{dist}(f,h) \leq \varepsilon$.

We will prove that size-s decision trees can be learned from queries with error ε in time $\mathrm{poly}(n, s, 1/\varepsilon)$. The proof is based on the Fourier 1-norm.

Definition 3.2. Let $f\colon \{\pm1\}^n \to \mathbb{R}$ or $f\colon \{0,1\}^n \to \mathbb{R}$, and let $p \geq 1$. The Fourier p-norm is defined by

$$\hat{\|}f\hat{\|}_p = \left(\sum_{S\subseteq[n]} |\widehat{f}(S)|^p\right)^{1/p}.$$

Lemma 3.3 (An extremely useful fact)**.** *If $f\colon \{0,1\}^n \to \{0,1\}$ is a conjunction of literals, then $\hat{\|}f\hat{\|}_1 = 1$.*

Proof. Without loss of generality, assume that f depends on every variable, i.e., $f(x) = 1 \iff x = x_*$ for some $x_* \in \{0,1\}^n$. Then for any $S \subseteq [n]$, we have

$$\widehat{f}(S) = \mathop{\mathbb{E}}_x[f(x)\cdot\chi_S(x)] = 2^{-n}\cdot\chi_S(x_*) = \pm 2^{-n}. \qquad \square$$

Corollary 3.4. *If $f\colon \{\pm1\}^n \to \{\pm1\}$ is a size-s decision tree, then $\hat{\|}f\hat{\|}_1 \leq s$.*

Proof. For each leaf u, there is some conjunction of literals f_u that indicates whether f reaches u on a given input x. Then $f(x) = \sum_u f_u \cdot \ell_u$ for some label $\ell_u \in \{\pm1\}$. Therefore,

$$\hat{\|}f\hat{\|}_1 \leq \sum_u |\ell_u| \cdot \hat{\|}f_u\hat{\|}_1 = s. \qquad \square$$

We will show that if a function has a bounded Fourier 1-norm, then its Fourier spectrum is concentrated on a relatively small set of coefficients, albeit not necessarily the low-degree coefficients.

Definition 3.5 (Concentration on a collection)**.** Let $f\colon \{\pm 1\}^n \to \mathbb{R}$ and let $\mathcal{F}$ be a collection of subsets of $[n]$. We say that (the Fourier spectrum of) f is *ε-concentrated on $\mathcal{F}$* if

$$\sum_{S \subseteq [n], S \notin \mathcal{F}} \widehat{f}(S)^2 \leq \varepsilon.$$

If f is $\{\pm 1\}$-valued, this is equivalent to the condition $\Pr_{S \sim \mathcal{S}_f}[S \notin \mathcal{F}] \leq \varepsilon$.

Lemma 3.6. *Let $f\colon \{\pm 1\}^n \to \mathbb{R}$. Then f is ε-concentrated on a collection of at most $\hat{\|} f \hat{\|}_1^2/\varepsilon$ coefficients.*

Proof. Let $\mathcal{F}$ be the set of $S \subseteq [n]$ such that $|\widehat{f}(S)| > \varepsilon / \hat{\|} f \hat{\|}_1$. Then f is ε-concentrated on $\mathcal{F}$, because

$$\sum_{S \subseteq [n], S \notin \mathcal{F}} \widehat{f}(S)^2 \leq \frac{\varepsilon}{\hat{\|} f \hat{\|}_1} \cdot \sum_{S \subseteq [n], S \notin \mathcal{F}} |\widehat{f}(S)| \leq \varepsilon.$$

Meanwhile, the cardinality of $\mathcal{F}$ is bounded, because

$$\hat{\|} f \hat{\|}_1 = \sum_{S \subseteq [n]} |\widehat{f}(S)| \geq \sum_{S \in \mathcal{F}} |\widehat{f}(S)| \geq |\mathcal{F}| \cdot \frac{\varepsilon}{\hat{\|} f \hat{\|}_1}. \qquad \square$$

Motivated by the calculations above, we wish to show that the class of functions that are concentrated on a small set of Fourier coefficients is learnable. To prove it, it will be convenient to encode sets using indicator vectors. So, for example, we use the notation $\widehat{f}(\gamma)$ and $\chi_\gamma(x)$ where $\gamma, x \in \{0,1\}^n$. Observe that $\chi_\gamma(x) = \chi_x(\gamma) = (-1)^{\sum_i x_i \gamma_i}$. We also use the following definition.

Definition 3.7 (The notation $W_\gamma[f]$)**.** Let $f\colon \{0,1\}^n \to \{\pm 1\}$ and $\gamma \in \{0,1\}^{\leq n}$. We define

$$W_\gamma[f] = \sum_{\beta \in \{0,1\}^{n - |\gamma|}} \widehat{f}(\gamma\beta)^2.$$

Equivalently, $W_\gamma[f]$ is the probability that a string drawn from the spectral sample $\mathcal{S}_f$ begins with γ.

The key to learning a function concentrated on a small set of Fourier coefficients is the Goldreich-Levin algorithm, which enables us to figure out *which* Fourier coefficients the function is concentrated on.

Theorem 3.8 (Goldreich-Levin algorithm)**.** *Suppose we are given query access to an unknown function $f\colon \{0,1\}^n \to \{\pm 1\}$, as well as a parameter $\theta \in (0,1]$. There is a randomized* $\mathrm{poly}(n/\theta)$*-time algorithm that makes queries to f and outputs a collection $\mathcal{F} \subseteq \{0,1\}^n$ such that with high probability:*

- *For every $\gamma \in \{0,1\}^n$, if $\widehat{f}(\gamma)^2 \geq \theta$, then $\gamma \in \mathcal{F}$.*
- *For every $\gamma \in \mathcal{F}$, we have $\widehat{f}(\gamma)^2 \geq \theta/4$.*

Proof. Algorithm: Let $\mathcal{F}_0 = \{0,1\}^0 = \{\text{empty string}\}$.

1. For $i = 1$ to n:
   (a) If $|\mathcal{F}_{i-1}| > 4/\theta$, abort.
   (b) For each $\gamma \in \mathcal{F}_{i-1}$ and each $b \in \{0,1\}$, estimate $W_{\gamma b}[f]$ to within $\pm\theta/4$. [We'll explain how to do this momentarily.]
   (c) If the estimate is at least $\theta/2$, then include γb in $\mathcal{F}_i$.
2. Output $\mathcal{F} = \mathcal{F}_n$.

We need to explain how to estimate $W_\gamma[f]$ for a given string γ. This is based on the following calculation:

$$\begin{aligned} W_\gamma[f] = \sum_\beta \widehat{f}(\gamma\beta)^2 = \sum_\beta \mathop{\mathbb{E}}_{xy}[f(xy) \cdot \chi_{\gamma\beta}(xy)]^2 &= \sum_\beta \mathop{\mathbb{E}}_{x,x',y,y'}[f(xy) \cdot f(x'y') \cdot \chi_{\gamma\beta}(xy) \cdot \chi_{\gamma\beta}(x'y')] \\ &= \mathop{\mathbb{E}}_{x,x',y}\left[2^{-|\beta|} \sum_{y'} f(xy) \cdot f(x'y') \cdot \chi_\gamma(x+x') \cdot \sum_\beta \chi_{y+y'}(\beta)\right] \\ &= \mathop{\mathbb{E}}_{x,x',y}[f(xy) \cdot f(x'y) \cdot \chi_\gamma(x+x')]. \end{aligned}$$

By Hoeffding's inequality, we can estimate $W_\gamma[f]$ to within $\pm\varepsilon$ with failure probability δ using $O(\log(1/\delta)/\varepsilon^2)$ queries to f, hence the algorithm runs in $\mathrm{poly}(n/\theta)$ time and we can assume that all the estimates are correct. Now let us prove that the algorithm succeeds, assuming all the estimates are correct.

By Parseval's theorem, the algorithm does not abort. Clearly, if $\gamma \in \mathcal{F}_n$, then $\widehat{f}(\gamma)^2 \geq \theta/4$. Conversely, let $\gamma \in \{0,1\}^n$ and suppose $\widehat{f}(\gamma)^2 \geq \theta$. Then $W_{\gamma_{1\dots i}}[f] \geq \theta$ for every $i \leq n$. Therefore, by induction, $\gamma_{1\dots i} \in \mathcal{F}_i$ for every $i \leq n$, and in particular $\gamma \in \mathcal{F}_n$. □

Corollary 3.9 (Kushilevitz-Mansour algorithm)**.** *The class of functions $f\colon \{\pm1\}^n \to \{\pm1\}$ such that f is ε-concentrated on a collection of at most M Fourier coefficients can be learned from queries with error 2ε in time $\mathrm{poly}(M, n, 1/\varepsilon)$.*

Proof. Suppose f is ε-concentrated on a collection $\mathcal{F}$ where $|\mathcal{F}| \leq M$. Let $\mathcal{F}' = \{S \subseteq [n] : \widehat{f}(S)^2 \geq \varepsilon/M\}$. Then f is (2ε)-concentrated on $\mathcal{F}'$:

$$\sum_{S : \widehat{f}(S)^2 < \varepsilon/M} \widehat{f}(S)^2 = \sum_{\substack{S \in \mathcal{F} \\ \widehat{f}(S)^2 < \varepsilon/M}} \widehat{f}(S)^2 + \sum_{\substack{S \notin \mathcal{F} \\ \widehat{f}(S)^2 < \varepsilon/M}} \widehat{f}(S)^2 \leq 2\varepsilon.$$

Therefore, run the Goldreich-Levin algorithm with $\theta = \varepsilon/M$, giving us a collection $\mathcal{F}''$ of subsets of $[n]$. We have $|\mathcal{F}''| \leq 4/\theta = 4M/\varepsilon$. Estimate $\widehat{f}(S)$ to within $\pm\sqrt{\varepsilon/|\mathcal{F}''|}$ for each $S \in \mathcal{F}''$; let c_S be the estimate. Hypothesis:

$$h(x) = \mathrm{sign}\left(\sum_{S \in \mathcal{F}''} c_S \cdot \chi_S(x)\right).$$

□

References

[EI22] Alexandros Eskenazis and Paata Ivanisvili. "Learning low-degree functions from a logarithmic number of random queries". In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*. 2022, 203–207. DOI: 10.1145/3519935.3519981.