Lecture 10

Lecture date: Sept 29 *Scribe: Sam Riesenfeld*

This lecture uses notation defined in the previous lecture; see the notes for Sept 27 for complete definitions and details. As a reminder, we summarize it here:

- $\mathcal{C}_n^k := \{f : \{-1,1\}^n \to \{-1,1\} \,:\, f \text{ depends on at most } k \text{ coords}\}$
- For $r > 0$, $\mathcal{C}_n^k(r) := \left\{f \in \mathcal{C}_n^k \,:\, \exists S,\ 0 < |S| \le r,\ \hat{f}(S) \ne 0\right\}$
- $\mathcal{C}_n^k(0) := \left\{f \in \mathcal{C}_n^k \,:\, \hat{f}(\emptyset) \ne 0\right\} = \left\{f \in \mathcal{C}_n^k \,:\, \mathbf{E}[f] \ne 0\right\}$

Let $\mathcal{C}_n$ be shorthand for $\mathcal{C}_n^n$.

We proved in the last lecture that $\mathcal{C}_n^k(0) \setminus \{-1,1\} \subseteq \mathcal{C}_n^k(\lceil \frac{2k}{3} \rceil)$ and that for any $f \in \mathcal{C}_n^k(r)$, we can find an influential variable in time $c_k n^r$.

In this lecture, we deal with functions like $\text{Parity}: \{-1,1\}^n \to \{-1,1\}$, defined as $\text{Parity}(x_1, x_2, \ldots, x_n) := x_1 \cdot x_2 \cdots x_n$, which is not contained in $\mathcal{C}_n(r)$ for any $r < n$.

To do this, we establish a bijection between $\{f : \{-1,1\}^n \to \{-1,1\}\}$ and $\{F : \{0,1\}^n \to \{0,1\}\}$ as follows:

$$\begin{aligned} F(X_1, X_2, \ldots, X_n) &:= \frac{1 - f((-1)^{X_1}, (-1)^{X_2}, \ldots, (-1)^{X_3})}{2} \\ &= \frac{1 - f(1-2X_1, 1-2X_2, \ldots, 1-2X_n)}{2}. \end{aligned}$$

In other words $F = \varphi f(\varphi^{-1} \times \cdots \times \varphi^{-1})$ where $\varphi : \{-1,1\} \to \{0,1\}$ is the homomorphism that takes -1 to 1, 1 to 0, and multiplication $(\cdot)$ to addition $(\oplus)$. (The notation $\varphi^{-1} \times \cdots \times \varphi^{-1}$ indicates the inverse of φ applied to each coordinate.)

Note that for $f \in \mathcal{C}_n^k$, the corresponding function F is a polynomial over the finite field $GF(2) = \mathbb{F}_2$ with degree $\deg_{\mathbb{F}_2} F \le k$. We define the class

$$\mathcal{P}_n^k(r) := \left\{f \in \mathcal{C}_n^k \,:\, \deg_{\mathbb{F}_2} F \le r \text{ for } F \text{ corresponding to } f\right\}.$$

Example 1 (Parity) *Let $f = x_1 \cdots x_n$ be the parity function. Then $F = x_1 \oplus x_2 \oplus \cdots \oplus x_n$. Notice that F is of low degree over $\mathbb{F}_2$ (f is in $\mathcal{P}_n(1)$), even though f has a high degree (n) in the Fourier basis!*

We recall that any function $F : \{0,1\}^n \to \{0,1\}$ can be written as a sum of multi-linear polynomials as follows:

$$F(X_1, \ldots, X_n) = \bigoplus_{S \subseteq [n]} \prod_{i \in S} X_i \prod_{i \notin S} (1 - X_i) F(1_{1 \in S}, 1_{2 \in S}, \ldots, 1_{n \in S}).$$

Claim 2 *The class $\mathcal{P}_n^k(r)$ can be learned in time n^{3r} (using Gaussian elimination) or $n^{\omega r}$, where $\omega = 2.367\ldots$ is the exponent for matrix multiplication.*

Proof: If $f \in \mathcal{P}_n^k(r)$, then $F = \bigoplus_{\substack{S\subseteq[n]\\|S|\leq r}} a_S \prod_{i\in S} X_i$. To determine f, it suffices to determine the coefficients $\{a_S\}$. Let $v(n,r)$ be the number of variables, that is, the number of coefficients. Then $v(n,r) = \binom{n}{r} + \binom{n}{r-1} + \cdots + \binom{n}{0} \leq O(n^r)$. Each sample of f gives a linear equation in the elements of $\{a_S\}$. Since solving a system of $\omega(n,r,\delta)$ that uniquely determine $v(n,r)$ variables takes time $O(n^{\omega r})$, the remaining part of the proof of Claim 2 is to answer the question: How many such equations are needed to determine these $v(n,r)$ variables? See Claim 3. $\square$

Claim 3 (Sub-claim of Claim 2) *Let $\omega(n,r,\delta) = 2^r(v(n,r) + \log\frac{1}{\delta})$. Given $\omega(n,r,\delta)$ samples, the coefficients in $\{a_S\}$ are determined uniquely except with probability δ.*

Proof: Let (a_S) be an assignment to the variables that satisfies all the equations given by the samples. Fix (b_S) such that $(b_S) \neq (a_S)$.

Exercise 4 (1pt; Prove by induction; This is called Schwarz-Zippel) *If x is chosen uniformly at random, then $\mathbf{P}[\bigoplus a_S \prod_{i\in S} x_i = \bigoplus b_S \prod_{i\in S} x_i] \leq 1 - 2^{-r}$.*

So an assignment (b_S), $(b_S) \neq (a_S)$, satisfies all equations given by the samples with probability at most $(1-2^{-r})^{\omega(n,r,\delta)}$. Thus

$$\begin{aligned}
&\mathbf{P}[\exists(b_S), (b_S) \neq (a_S),\ (b_S) \text{ satisfies all equations}]\\
&\leq \mathbf{E}[\#(b_S) \neq (a_S) \text{ that satisfy all equations}]\\
&\leq 2^{v(n,r)}(1-2^{-r})^{\omega(n,r,\delta)}\\
&\leq \delta.
\end{aligned}$$

$\square$

Lemma 5 ("Lemma 1" in class) *For every r, $1 \leq r \leq k$, the following holds: every $f \in \mathcal{C}_n^k$ belongs to one of the following families:*

- $\{-1, 1\}$ (1)
- $\mathcal{C}_n^k(0) \setminus \{-1,1\} \subseteq \mathcal{C}_n^k(\lceil \frac{2k}{3} \rceil)$ (2)
- $\mathcal{C}_n^k(r)$ (3)
- $\mathcal{P}_n^k(k-r)$. (4)

Claim 6 *There is a $c_k n^{\alpha k + O(1)} + c_k \log\frac{1}{\delta}$ learning algorithm for $\mathcal{C}_n^k$, with $\alpha = \frac{\omega}{\omega+1}$.*

Proof (Of Claim 6): Note that $\alpha \geq \frac{2}{3}$. We check that every $f \in \mathcal{C}_n^k$ falls into a class that we can learn in the time listed in Lemma 5. (We ignore dependencies on δ here.)

- Class (1): obvious.
- Class (2): We showed last lecture that $\mathcal{C}_n^k(0)$ can be learned in time $O(n^{\frac{2k}{3}})$.
- Class (3) for $r = \alpha k$: We showed last lecture that $\mathcal{C}_n^k(r)$ can be learned in time $O(n^r)$, which is $O(n^{\alpha k})$ for $r = \alpha k$. If $f \notin \mathcal{C}_n^k(r)$ for $r = \alpha k$, then, as we shall see in the proof of Lemma 5, the corresponding function F is in $\mathcal{P}_n^k(k-r)$.
- Class (4) for $r = \alpha k$: We showed during this lecture that $\mathcal{P}_n^k(k-r)$ can be learned in time $O(n^{\omega(k-r)})$, which is $O(n^{\alpha k})$ for $r = \alpha k$.

□

Proof (Of Lemma 5): We need to show that if $f \notin \mathcal{C}_n^k(r)$ and $\mathbf{E} f = 0$, then $F \in \mathcal{P}_n^k(k-r)$. For this proof, we assume that $n = k$, which is not constraining since n is not relevant in the statement.

Since $f \notin \mathcal{C}_k(r)$, we can write

$$f = \sum_{|S|>r} \hat{f}(S) \prod_{i \in S} x_i.$$

Let $g = f \prod_{i=1}^k x_i$. Notice that g has only low Fourier coefficients and can be written as

$$g = \sum_{|S|<k-r} \hat{f}([k] \setminus S) \prod_{i \in S} x_i.$$

Now we observe that the corresponding function G is

$$G = F \oplus X_1 \oplus \cdots \oplus X_n,$$

and that $\deg_{\mathbb{F}_2} F = \deg_{\mathbb{F}_2} G$, since adding $X_1 \oplus \cdots \oplus X_n$ does not change the degree in $\mathbb{F}_2$. The proof concludes with Lemma 7. □

Lemma 7 ("Lemma 2" in class) *For every* $g :\{-1,1\}^k \rightarrow \{-1,1\}$ *and corresponding* $G :\{0,1\}^k \rightarrow \{0,1\}$, *that is,* $G = \varphi g(\varphi^{-1} \times \cdots \times \varphi^{-1})$, $\deg_{\mathbb{R}} g \geq \deg_{\mathbb{F}_2} G$.

Proof: Let us imagine for a moment that $\varphi :\{-1,1\} \rightarrow \{0,1\}$, $\varphi(x) = \frac{1-x}{2}$, is a map into $\mathbb{R}$. Since φ is linear, $(\varphi^{-1} \times \cdots \times \varphi^{-1})$ is linear in each of the k coordinates. Since the maps are linear, they don't change the degree, and thus $\deg_{\mathbb{R}} g = \deg_{\mathbb{R}} G$. It therefore suffices to show that $\deg_{\mathbb{R}} G \geq \deg_{\mathbb{F}_2} G$. We recall the trivial way of writing a $0,1$-valued function as a polynomial. Over $\mathbb{R}$, we have the equality

$$G(X_1, \ldots X_k) = \sum_{z \in \{0,1\}^k} G(z) \prod_{i\,:\,z_i=1} x_i \prod_{i\,:\,z_i=0} (1-x_i). \tag{5}$$

Taking equation (5) mod 2 can only reduce the degree. Assuming the next lemma, i.e. that there is a unique multi-linear representation of G over $\mathbb{R}$, we are done. □

Lemma 8 *G has a unique multi-linear representation over $\mathbb{R}$.*

Proof: Proof by induction on the dimension k. If $k = 1$, then G is a linear function from $\mathbb{R}$ to $\mathbb{R}$ and so we can write $G(x) = ax + b$ for fixed constants a and b, which gives a unique linear representation. Suppose the lemma holds for all $k < n$. Let $k = n + 1$. If G is multi-linear, then we can write $G(x_1, \ldots, x_n, x_{n+1}) = F_1(x_1, \ldots, x_n)x_{n+1} + F_2(x_1, \ldots, x_n)$, where F_1 and F_2 are multi-linear. If G had different multi-linear representations, then so would F_1 or F_2 (or both). $\square$