

Lecture 7 : March 1, 2007

Lecturer: Rocco A. Servedio

Scribe: Daniel L.C. Mack

1 Today's Lecture

- Fourier Representation of Boolean Functions:
 1. Basic Definition
 2. Parity functions as a basis
 3. Examples
- Learning with Fourier Representations:
 1. Estimating Fourier coefficients
 2. Fourier Concentration
 3. Low Degree Algorithm

2 Fourier Representation of Real-Valued Functions

Let the domain be $\{-1, 1\}^n$, where -1 represents true and 1 represents false. We will consider functions that map $\{-1, 1\}^n$ to the reals, $\{-1, 1\}^n \rightarrow \mathbb{R}$. Any such real function is defined by 2^n real numbers. Boolean functions are the special case that all function values happen to lie in $\{-1, 1\}$.

Let V be the vector space that contains all functions $\{-1, 1\}^n \rightarrow \mathbb{R}$. Clearly V is closed under addition of functions and scalar multiplication.

As an example, let $n = 2$ and $f(1, 1) = 3$, $f(1, -1) = 5$, $f(-1, 1) = -2$, and $f(-1, -1) = 4$. Then f can be represented as the vector $f = (3, 5, -2, 4)$. This is a

nice representation if we want to use the standard basis. This vector could be broken down into the components of the standard basis as:

$$f = 3(1, 0, 0, 0) + 5(0, 1, 0, 0) - 2(0, 0, 1, 0) + 4(0, 0, 0, 1)$$

Each of the 2^n standard basis elements corresponds to a function δ_z defined by

$$\delta_z(x) = 1 \text{ if } x = z, \quad \delta_z(x) = 0 \text{ if } x \neq z.$$

This basis allows us to map each distinct input from the function to its corresponding coordinate and scale it appropriately.

3 The Parity Basis

This standard basis is not the ideal basis for the vector space V because it ignores the structure of the Boolean hypercube that is the input domain for the functions.

Instead we will consider as a basis the set of all 2^n parity functions. The parity function is defined for a set $S \subseteq 1, 2, \dots, n$. For $x \in \{-1, 1\}^n$ the function $PAR_S(x)$ is -1 when an odd number of x_i 's are set to -1 and 1 if an even number is set to -1. For example, for the set $S = \{2, 4, 5\}$ the function $PAR_S(1, 1, -1, -1, -1) = 1$.

There is a more concise way to represent this:

$$PAR_S(x) = \prod_{i \in S} x_i.$$

We will represent $PAR_S(x)$ as $\chi_S(x)$ from now on. Note that for $S = \emptyset$ (the empty set) the parity function $\chi_\emptyset(x)$ is the identically-1 function.

We define an inner product over V :

Definition 1 Given $f, g \in V$, we define

$$\langle f, g \rangle \stackrel{\text{def}}{=} \frac{1}{2^n} \sum_{x \in \{-1, 1\}^n} f(x)g(x) = \mathbf{E}[f(x)g(x)]$$

where the expectation is taken w.r.t. the uniform distribution over $\{-1, 1\}^n$.

The corresponding norm over V is $\|f\| \stackrel{\text{def}}{=} \sqrt{\langle f, f \rangle} = \sqrt{\mathbf{E}[f(x)^2]}$.

3.1 Properties of Parity Functions

Here are some useful properties of parity functions:

1. Each parity has unit norm: for any S , $\chi_S(x)^2 = 1$ for all x and hence $\|\chi_S\| = 1$.
2. The product of two parity functions is another parity function with a set composed of the symmetric difference.

$$\chi_T(x)\chi_S(x) = \chi_{S\Delta T}(x).$$

For example, $\chi_S(x) = x_1x_2x_3$ and $\chi_T(x) = x_2x_3x_4$, become:

$$\chi_S\chi_T = x_1x_2^2x_3^2x_4 = x_1x_4$$

3. $\mathbf{E}[\chi_\emptyset(x)] = 1$ and $\mathbf{E}[\chi_S(x)] = 0$ for all $S \neq \emptyset$.
4. The functions χ_S , for $S \subset [n]$, are orthogonal to one another. For $S \neq T$, the inner product of two parity functions

$$\langle \chi_T, \chi_S \rangle = \mathbf{E}[\chi_T(s)\chi_S(x)]$$

simplifies (using the second property from above) to

$$\mathbf{E}[\chi_{T\Delta S}(x)].$$

Since $S \neq T$, the symmetric difference is nonempty and by (3) above the expectation is zero.

5. The 2^n functions χ_S , $S \subseteq [n]$, are linearly independent (this follows easily from orthogonality).

These properties show that parities form an orthonormal basis for V . Consequently any $f \in V$ can be uniquely expressed as a linear combination of χ_S , ($S \in [n]$): i.e. we have a unique expression

$$f(x) = \sum_{S \in [n]} \hat{f}(S)\chi_S(x).$$

We say that $\hat{f}(S)$ is the S -th Fourier coefficient of the function f . The value of this coefficient is $\hat{f}(S) = \langle f, \chi_S \rangle = \mathbf{E}[f \cdot \chi_S]$: note that we have

$$\langle f, \chi_S \rangle = \left\langle \sum_{T \in [n]} \hat{f}(T)\chi_T(x), \chi_S(x) \right\rangle = \sum_{T \in [n]} \hat{f}(T)\langle \chi_T(x), \chi_S(x) \rangle = \hat{f}(S)$$

since the inner product $\langle \chi_T, \chi_S \rangle$ is zero unless $S = T$ in which case it is 1.

3.2 Fourier representation and Boolean functions

Now consider the case where f is a Boolean function (with range $\{-1, 1\}$). We have

$$\hat{f}(S) = \mathbf{E}[f \cdot \chi_S] = \Pr[f(x) = \chi_S(x)] - \Pr[f(x) \neq \chi_S(x)]$$

Letting p denote $\Pr[f(x) = \chi_S(x)]$, this is equal to $p - (1 - p) = 2p - 1$. We thus may view the Fourier coefficient of a Boolean function as a measure of how often the function agrees with the corresponding parity. If p is 1 (and the Fourier coefficient is 1) then they always agree; if $p = 1/2$ (and the Fourier coefficient is 0) then they agree half the time, i.e. they are uncorrelated; and if $p = 0$ (and the Fourier coefficient is -1) then they always disagree.

It is clear from the above that $\hat{f}(S) \in [-1, 1]$ for any Boolean function, but the following result implies something much stronger:

Theorem 1 [*Plancherel's identity*] For any $f, g : \{-1, 1\}^n \rightarrow \mathbb{R}$ we have:

$$\sum_{S \subseteq [n]} \hat{f}(S) \hat{g}(S) = \frac{1}{2^n} \sum_{x \in \{-1, 1\}^n} f(x) g(x).$$

Proof:

$$\frac{1}{2^n} \sum_{x \in \{-1, 1\}^n} f(x) g(x) = E[f \cdot g] = E\left[\left(\sum_S \hat{f}(S) \chi_S(x)\right) \left(\sum_T \hat{g}(T) \chi_T(x)\right)\right] =$$

$$E\left[\sum_S \sum_T \hat{f}(S) \hat{g}(T) \chi_S(x) \chi_T(x)\right] = \sum_S \sum_T \hat{f}(S) \hat{g}(T) E[\chi_S(x) \chi_T(x)]$$

When $S = T$, the expectation is equal to 1, otherwise the expectation is zero and the terms disappear. This means that the above double sum simplifies to

$$\sum_{S \subseteq [n]} \hat{f}(S)^2.$$

■

An important corollary is that $\sum_{S \subseteq [n]} \hat{f}(S)^2$ is exactly 1 for any Boolean function.

As an example, let us consider the Fourier spectrum of $\text{AND}(x_1, x_2, \dots, x_k)$ where the output is -1 if all variables are equal to -1 and 1 otherwise. Let $f(x)$ be:

$$f(x) = \frac{1 - x_1}{2} \cdot \frac{1 - x_2}{2} \cdot \dots \cdot \frac{1 - x_k}{2}$$

which can equivalently be represented in the parity basis as:

$$f(x) = \sum_{s \in [k]} \frac{-1^{|S|}}{2^k} \chi_S(x)$$

The fourier coefficients are $\frac{-1^{|S|}}{2^k}$ and $\text{AND}(x_1, x_2, \dots, x_n) = 1 - 2f(x)$. Thus the coefficients are $\hat{A\hat{N}D}(S) = 1 - \frac{2}{2^k}$.

4 Learning

The intuitive connection between learning and Fourier representations is that every boolean function's Fourier coefficients have sum of squares equal to 1; think of this as the "power" that the function has on various parity basis functions. If we can find or approximate most of this Fourier weight ($1 - \epsilon$ of it), then intuitively we should be able to ϵ -approximate the function. Next we will discuss approximating Fourier coefficients.

4.1 Approximation

Recall that in our model we have an example oracle $EX(f)$ where we have a uniform distribution over the examples. Given such an oracle and a particular set such as $S = \{1, 2, 3\}$, can we estimate $\hat{f}(S)$? The answer is yes.

Lemma 1 *Fix $S \subseteq [n]$. Given S , $\delta > 0, \gamma > 0$ and given the example oracle, we can estimate and obtain value C_S such that $|C_S - \hat{f}(S)| \leq \gamma$ with probability $\geq 1 - \delta$ using $O(\frac{\log(\frac{1}{\delta})}{\gamma^2})$ calls to the oracle.*

Proof: A standard application of Chernoff bounds shows that the claimed number of uniform random examples drawn from $EX(f)$ gives an estimate of $\mathbf{E}[f(x)\chi_S(x)]$

that is accurate to within an additive $\pm\gamma$ with probability $1 - \delta$. (This uses the fact that the value of $f(x)\chi_S(x)$, for every x , is ± 1 .) ■

The above lemma says that any *particular* Fourier coefficient can be efficiently estimated; but how about *finding* an unknown Fourier coefficient whose magnitude is large? More concretely, suppose that $f : \{-1, 1\}^n \rightarrow \{-1, 1\}$ is a Boolean function for which some $S \subseteq [n]$ satisfies $|\hat{f}(S)| \geq \frac{9}{10}$. Given access to a random example oracle $EX(f)$, is it possible to find S in $\text{poly}(n)$ time? The answer is not known; at this point no $\text{poly}(n)$ -time algorithm is known to exist. (An easy approach of drawing $\text{poly}(n)$ examples and then estimating *all* 2^n Fourier coefficients using this fixed sample is easily seen to work, but this takes 2^n time steps.)

What can we do? We can estimate all Fourier coefficients corresponding to sets of variables of size $1, 2, \dots, d$ in time $n^{O(d)}$. So intuitively, if f has almost all of its Fourier weight on these coefficients, it should be possible to learn f in $\text{poly}(n^d)$ time. We now make this intuition precise.

Definition 2 Let $f: \{-1, 1\}^n \rightarrow \mathbb{R}$. We say f has $\alpha(\epsilon, n)$ Fourier Concentration if

$$\sum_{|S| > \alpha(\epsilon, n), S \subseteq [n]} \hat{f}(S)^2 \leq \epsilon.$$

If f is boolean then the above is equivalent to:

$$\sum_{|S| \leq \alpha(\epsilon, n), S \subseteq [n]} \hat{f}(S)^2 \geq 1 - \epsilon.$$

4.2 Low Degree Algorithm

We now describe the Low Degree Algorithm for a function f with Fourier Concentration $\alpha(\epsilon, n)$. The algorithm takes values τ, δ as input parameters:

1. Let $d = \alpha(\epsilon, n)$, Let $m = O(\frac{n^d}{\tau} \log(\frac{n^d}{\delta}))$. Make m draws from the example oracle $EX(f)$. Use the sample to obtain C_S , an estimate of $\hat{f}(S)$, for all S with all $|S| \geq d$
2. Output $h(x) = \sum_{|S| \leq d} C_S \chi_S(x)$. Note that $h : \{-1, 1\}^n \rightarrow \mathbb{R}$, i.e. h is a real-valued hypothesis.

Theorem 2 If f has $\alpha(\epsilon, n)$ Fourier Concentration then with probability $\geq 1 - \delta$ we have $\mathbf{E}[(f(x) - h(x))^2] \leq \epsilon + \tau$.

Proof: Let $\gamma = \sqrt{\frac{\tau}{n^d}}$ so $(\frac{1}{\gamma^2} = \frac{n^d}{\tau})$.

The earlier lemma implies that for each set S of size m , the sample gives us C_S with $|\hat{f}(S) - C_S| > \gamma$ with probability $\leq \frac{\delta}{n^d}$.

A union bound over all sets S gives us that for $|S| \leq d$ with probability $\leq 1 - \delta$ every C_S satisfies $|\hat{f}(S) - C_S| \leq \gamma$. ■

To Be Continued in our next exciting lecture, stay tuned...