PRINCETON UNIV. F'07 COS 597D: THINKING LIKE A THEORIST

Lecture 22: Fourier Analysis Continued

Lecturer: *Sanjeev Arora* Scribe:*Hart Montgomery*

Let's begin by recapping a bit from the last lecture. Recall that we are primarily interested in generating a Fourier basis for $f : \{-1, 1\}^n \rightarrow \Re$, where $\Re$ is defined as a vector on $\Re^{2^n}$. We want an orthonormal basis for $\Re^{2^n}$ where $\{\chi_s\}_{s \subseteq \{1,2,\ldots,n\}}$ such that $\chi_s(x) = \prod_{i \in s} x_i$. Additionally, remember that we defined

$$f(x) = \sum_s \widehat{f}(s) \chi_s(x)$$

as our general Fourier approximation. Now, before we go further, remember the definition

$$f \circ g = E_x [f(x) g(x)]$$

and also that we proved Parseval's Identity in the last lecture for functions of this type, which states that

$$\sum \widehat{f}(s)^2 = ||f||_2^2 = 1$$

These facts will be useful to us later in this lecture. Last time, our main result was to prove that if $I(f) \leq 1 + \epsilon$, then f has ℓ_2 diestance $O(\epsilon)$ to a coordinate function, where I denoted the influence of f. Since this is well explained and defined in the previous lecture, we omit further discussion of this here.

This is a nice result, but in an ideal world we would like something more general. It turns out that we can, in fact, generalize this statement that we proved earlier. Today we are going to show a result first proven by Friedgut in '97, namely that

THEOREM 1
If f is balanced, boolean, and $I(f) = k$, then f has ℓ_2 distance of $\leq \epsilon$ to a boolean function of at most $2^{O\left(\frac{k}{\epsilon}\right)}$ variables for all $\epsilon > 0$.

Before we go any further, we mention some terminology common to the field (although it is not used in Friedgut's paper itself). For whatever reason, people like to nickname groups of boolean functions by forms of government based upon how many variables dictate the function and how many people are involved in the government. Thus, a coordinate function is called a dictatorship, and a function of a small number of variables is called a junta.

With this aside, let's continue on. The intuition for this argument should be fairly simple if you understand the previous lecture, so let's move on to the proof itself, which will take up the majority of this lecture.

Suppose that we define a constant $d = \frac{2k}{\epsilon}$. Recall that

$$f = \sum \widehat{f}(s) \chi_s$$

and also that

$$\sum |s|\widehat{f}(s) = k$$

Now let's consider the ℓ_2 norm of the high-degree variables (those with some degree greater than d) in f. With the previous statements in mind, it should be easy enough then to see that

$$\sum_{|s|>d} \widehat{f}(s)^2 \leq \frac{k}{d} = \frac{\epsilon}{2}$$

which means that the high degree terms are insignificant, and we can dismiss them without loss of generality due to our bound in the theorem. While this is a nice result, we still have significant work to do in the overall proof. We just showed that we can consider f to be a degree d polynomial. However, f could still be a function of all n variables. Next, we need to outline a solution to deal with this problem. Suppose that we define a function J such that

$$J = \{j : Inf_j(f) \geq 100^{-d}\}$$

so that $|J| \leq 100^d k e^{\frac{k}{\epsilon}}$. Now let's suppose that we ignore the variables outside of J. If we could do this, then we could use the following function to satisfy our theorem:

$$g(x) = sign\left(\sum_{s:s\subseteq d, s\subseteq J} \widehat{f}(s)\chi_s\right)$$

However, we do not yet know that the terms that we just threw away by not including terms that are not in J are negligible yet, so we can only label this a candidate close function at this point. However, it turns out that we can prove that the terms outside of J are in fact negligible. To do this, let's consider the terms that we threw away in this step. We know that the ℓ_2 norms of these terms must correspond to

$$\sum_{s:|s|\leq d, s\cap J^c\neq 0} \widehat{f}(s)\chi_s$$

Thus, to finish the proof we must bound this ℓ_2-norm by $\frac{\epsilon}{2}$. Our characterization of sensitivity last time shows that for all $j \in \overline{J}$,

$$\sum_{s:j\in s} \widehat{f}(s)^2 < 100^{-d}$$

Now let's do some calculations. We know that

$$\sum_{s\cap J^c\neq\emptyset} \widehat{f}(s)^2 \leq \sum_{s\cap J^c\neq\emptyset} |s\cap J^c|\widehat{f}(s)^2 \leq 4^{d-1}\sum_{s\cap J^c\neq\emptyset} |s\cap J^c|\frac{1}{4^{d-1}}\widehat{f}(s)^2$$

and it is also easy to see that

$$4^{d-1}\sum_{s\cap J^c\neq\emptyset} |s\cap J^c|\frac{1}{4^{d-1}}\widehat{f}(s)^2 \leq 4^{d-1}\sum_{s\cap J^c\neq\emptyset} |s\cap J^c|\frac{1}{4^{|s|-1}}\widehat{f}(s)^2$$

It looks as though these calculations take us nowhere. However, it turns out that the function

$$\sum_S \frac{1}{4^{|s|-1}} \widehat{f}(s)^2$$

is actually quite analyzable, and that many theoretical computer science applications of Fourier analysis boil down to this expression (who knew?). In order to exploit this, we are going to have to change gears and take things from an entirely different perspective. Thus, we turn to the work of Nelson, Bonamie, Beckner, etc., which allows us to say something meaningful about expressions of the form

$$\sum \rho^{|s|-1} \widehat{f}(s)^2$$

where $\rho < 1$. This expression arises when we analyse *noise operators* on functions. Let a noise operator $T_\delta(f)$ be an operator on functions from $\{-1,1\}^n$ to $\Re$ where the output is a "smoothed" version of f. The value of $T_\delta(f)$ is the expectated value of f on all inputs where each bit is flipped with probability δ.

$$T_\delta(f)(x) \equiv E_\eta[f(x * \eta)]$$

where $\eta \in \{-1,1\}^n$ is a vector in which each coordinate is independently chosen to be -1 with probability δ. (Thus it has about $n\delta$ coordinates that are -1.) Thus, the Hamming ball of radius $n\delta$ around the true output of f gets averaged, which is much like a smoothing operation seen in many distinct applications (using Fourier analysis for smoothing is useful for image processing and compression, a whole different topic in a very different field). This diversion aside, let's do some examples to get a feel for $T_\delta(f)$. We know that if $f = x_i$, then $T_\delta(f) = x_i(1-2\delta)$, which, in terms of noise and smoothing, is the effect of a small attenuation. If we let $f = \chi_s$, then we can see that $T_\delta(f)$ must be mapped to $(1-2\delta)^{|s|}\chi_s$. From the previous line we conclude using linearity of expectations that the fourier expansion of $T_\delta(f)$ is

$$T_\delta(f) = \sum_s (1-2\delta)^{|s|} \widehat{f}(s)\chi_s$$

and from this,

$$|T_\delta(f)|_2^2 = \sum_s (1-2\delta)^{2|s|} \widehat{f}(s)$$

The Beckner-Bonamie inequality concerns l_p norms of f and $T_\delta(f)$. We define

$$|f|_p = (E_x[f(x)^p])^{\frac{1}{p}}.$$

It is an elementary fact about norms that for $1 \leq p < q$,

$$|f|_p < |f|_q$$

At this point, you might be wondering if it is possible to find a constant c such that

$$|f|_q \leq c - |f|_p$$

Unfortunately, the answer in general is no, because c must be dependent upon f. However, there are some special cases that we can exploit, and Beckner proves one of them for us in the following theorem, which ties together what we have mentioned previously quite nicely.

Theorem 2 (hypercontractivity estimate)
$|T_\delta(f)|_q \leq ||f||_p$ *if* $1-\delta < \sqrt{\frac{p-1}{q-1}}$.

The proof takes some work and can be found in O'Donnell's lecture notes.

With this theorem in mind, we can go back to the calculations in the original problem and make some progress. Recall the last expression that we encountered. Applying the previous theorem to it, we can see that

$$4^{d-1} \sum_{s \cap J^c \neq \emptyset} |s \cap J^c| \frac{1}{4^{|s|-1}} \widehat{f}(s)^2 \leq 4^{d-1} \sum_{i \in \overline{J}} \sum_{s: i \in s} \left(\frac{1}{4}\right)^{|s|-1} \widehat{f}(s)^2$$

Now suppose that we add another definition, which we explain below:

$$D_i \equiv \sum_{i \in s} \widehat{f}(s) \chi_{s \setminus \xi_i}$$

where ξ_i is one of the n variables of f. Thus, D_i is just one of the possible Fourier series of f if we can only use $n-1$ variables to approximate f instead of n. With this in mind, it should be easy enough to see that

$$\sum_{s: i \in s} \left(\frac{1}{4}\right)^{|s|-1} \widehat{f}(s)^2$$

is just the ℓ_2 norm of $T_\delta(D_i(f))$. Therefore, we can say that

$$4^{d-1} \sum_{i \in \overline{J}} \sum_{s: i \in s} \left(\frac{1}{4}\right)^{|s|-1} \widehat{f}(s)^2 \leq 4^{d-1} \sum_{i \in \overline{J}} ||D_i(f)||_q^2$$

It should be pretty clear where we want to go from here. All we need to do is insert the bound where, for all $j \in \overline{J}$,

$$\sum_{s: j \in s} \widehat{f}(s)^2 < 100^{-d}$$

into the equation for the summation, and it's clear that this term blows away the 4^{d-1} term, which allows us to bound the ℓ_2 norm of the terms not in J to a very small number. With this bound, we know that our candidate close function which was mentioned earlier must actually be close, which is all we needed to show to prove the theorem that we proposed earlier.

In the last 15-20 min, we briefly discussed the use of the noise operator in Hastad's 1997 paper on inapproximability results (available online from the ACM portal). Namely, we discussed a test for being a coordinate function.

Suppose that we have a function $f : \{-1,1\}^n \to \{-1,1\}$. If f is a coordinate function, then $f(x)f(y) = f(x*y)$ for all choices of x and y. Then a test for the coordinate function proceeds as follows: pick $x, y \in \{-1,1\}^n$, a noise vector η, and check if $f(x)f(y) = f(x*y*\eta)$. If f is a coordinate function, the test succeeds with probability $1-\delta$. However, what is more useful is to show that if the test succeeds with high probability, then f is close to a coordinate function (or something like that).

The analysis proceeds by noticing that the error probability of the test can be expressed as

$$E\left[f\left(x\right)f\left(y\right)f\left(x*y*n\right)\right]\sim\sum_{s}\widehat{f}\left(s\right)^{3}\left(1-2\delta\right)^{|s|}$$

from which the result follows by not a very difficult calculation.

Then we discussed an analog of the Hastad test using two bits: if we pick x,η as above and then check if $f\left(x\right)=f\left(x*\eta\right)$. This has proved useful in recent PCP constructions. But proving its correctness requires a lemma of Bourgain that also uses the Beckner-Bonamie inequality. The final thing that was briefly described was the "Majority is Stablest" theorem.