6.893 Approximability of Optimization Problems	October 27, 1999
Lecture 13	
Lecturer: Madhu Sudan	*Scribe: Ryan O'Donnell*

1 Linearity Testing Over {0, 1} With Fourier Analysis

In this lecture we look at an alternative view of linearity testing, introduced in some papers of Bellare, Coppersmith, Håstad, Kiwi, and Sudan.

Previously we viewed our code functions $f : \{0,1\}^m \to \mathbf{Z}_2$ as taking values in the additive group of two elements; we now let the range be $\{+1, -1\}$ instead, with $+1$ in place of 0, and -1 in place of 1. So x in our old range corresponds to $(-1)^x$ in our new range, and addition corresponds to multiplication.

This switch to $\{0,1\}$ is really only syntactic, but it can often make our work much cleaner. For an illustration of this, try working out the following exercise, using an additive $\{0,1\}$ representation first, and then a multiplicative $\{+1,-1\}$ representation:

Exercise 1 *Suppose we have a biased coin which outputs 0 with probability $1-\epsilon$ and 1 with probability ϵ. We wish to simulate an unbiased coin. One idea is to flip k coins and then output the parity of the k 0/1 results. Show that with k large this gives an essentially unbiased result.*

So we now have functions $f : \{0,1\}^m \to \{+1,-1\}$, which pass the linearity test exactly when $f(x+y) = f(x)f(y)$. What can we say regarding the degree to which passing the linearity test is correlated with being close to a linear function? The key idea is to view our functions as taking values in $\mathbf{R}$; then our functions f are really vectors in a 2^m-dimensional vector space of functions. If we now pick a "nice" basis for this vector space, we can do some analysis which will let us link the linearity test with the property of being close to a linear function.

Recall some definitions from linear algebra:

Definition 2 *A* (real) inner product space *is a vector space V over the field $\mathbf{R}$ (i.e., a set of vectors which we can add and subtract, and which we can multiply by real scalars), along with an inner product $\langle , \rangle : V \times V \to \mathbf{R}$ satisfying:*

- $\langle a, b\rangle = \langle b, a\rangle \ \forall a, b \in V$
- $\langle a + b, c\rangle = \langle a, c\rangle + \langle b, c\rangle \ \forall a, b, c \in V$
- $\langle \lambda a, b\rangle = \lambda\langle a, b\rangle \ \forall \lambda \in \mathbf{R}, \forall a, b \in V$
- $\langle a, a\rangle \geq 0 \ \forall a \in V$ *with equality iff* $a = 0$

Definition 3 *In an inner product space $(V, \langle,\rangle)$, a set $\{u_1, \ldots, u_n\}$ is an* orthonormal basis *if every vector $v \in V$ can be written as a linear combination of the u_i's, and further $\langle u_i, u_j\rangle = 0\ \forall i \neq j$, $\langle u_i, u_i\rangle = 1\ \forall i$.*

We also have the following elementary facts from Fourier Analysis:

Proposition 4 *Let $\{u_1, \ldots, u_n\}$ be an orthonormal basis for $(V, \langle,\rangle)$. If $x \in V$:*

- *(representation via Fourier coefficients)* $x = \sum_{i=1}^{n} \langle x, u_i\rangle u_i$
- *(Parseval's identity)* $\langle x, x\rangle = \sum_{i=1}^{n} \langle x, u_i\rangle^2$

The inner product space we will study will have V as the set of functions $\{0,1\}^m \to \mathbf{R}$, and $\langle f, g\rangle = \frac{1}{2^m} \sum_{x\in\{0,1\}^m} f(x)g(x)$. We added the normalization factor of $\frac{1}{2^m}$ so that the particular functions we're interested in — those that take values in $\{+1, -1\}$ — will all have unit length; for notice that if $f(x)$ is always ± 1 then $f(x)^2$ is always 1, so $\langle f, f\rangle = \frac{1}{2^m} \sum_{x\in\{0,1\}^m} f(x)^2 = 1$.

Definition 5 *$\ell_{\mathrm{a}} \in V$ is the function given by $\ell_a(x) = (-1)^{\sum_{i=1}^{m} a_i x_i}$, where a is a vector of coefficients in $\{0,1\}^m$.*

Recalling our transformation from $\{0, 1\}$ and $+$ to $\{+1, -1\}$ and $\times$, we see that our linear functions in the new setting are exactly the functions ℓ_a. These functions also have a very nice property that they form an orthonormal basis for V. We'll prove this shortly.

In our inner product space, if we have two code functions f, g mapping $\{0,1\}^m$ into $\{+1, -1\}$, what interpretation does $\langle f, g\rangle$ have?

Proposition 6 $\langle f, g\rangle = 1 - 2\mathrm{Pr}_x[f(x) \neq g(x)]$

Proof

Note that $f(x) = g(x) \Rightarrow f(x)g(x) = +1$, whereas $f(x) \neq g(x) \Rightarrow f(x)g(x) = -1$. Thus:

$$\begin{aligned}
\langle f, g\rangle &= \frac{1}{2^m} \sum_{x\in\{0,1\}m} f(x)g(x) \\
&= \mathrm{Ex}_x[f(x)g(x)] \\
&= (+1)\cdot \mathrm{Pr}_x[f(x) = g(x)] + (-1)\cdot \mathrm{Pr}_x[f(x) \neq g(x)] \\
&= 1 - 2\mathrm{Pr}_x[f(x) \neq g(x)]
\end{aligned}$$

■

As a consequence, we get:

Proposition 7 *The set $\{\ell_a\}_{a\in\{0,1\}^m}$ forms an orthonormal basis for our inner product space.*

Proof

First we check that $\langle \ell_a, \ell_b \rangle = \delta_{ab}$. We have $\langle \ell_a, \ell_a \rangle = 1$ for all a since ℓ_a takes values in $\{+1, -1\}$. We also claim that for $a \neq b$, $\Pr[\ell_a(x) \neq \ell_b(x)] = 1/2$. The probability is at most 1/2 because of the Schwarz-Zippel result: we have distinct degree 1 functions over a field of 2 elements. Since this holds for any two linear functions, we can get the conclusion from the fact that both $\Pr[\ell_a(x) \neq \ell_b(x)] \leq 1/2$ and $\Pr[(-\ell_a)(x) \neq \ell_b(x)] \leq 1/2$. It follows now that for $a \neq b$, $\langle \ell_a, \ell_b \rangle = 0$.

Now that we have the desired orthogonality conditions, it follows immediately that $\{\ell_a\}_{a \in \{0,1\}^m}$ is in fact a basis.

■

We've shown that we can tell how far apart any two functions are by looking at their inner product. Recall that our desire is that f passing the linearity test with high probability should imply that f is close to a linear function. Let's formalize the consequent:

Definition 8 *For a given function f, define $\epsilon(f)$ to be $min_a\{\Pr_x[f(x) \neq \ell_a(x)]\}$.*

Note that by the previous proposition, $\epsilon(f)$ is just $\min_a\{\frac{1-\langle f, \ell_a \rangle}{2}\}$, or equivalently, $\frac{1-\max_a\{\langle f, \ell_a \rangle\}}{2}$. So to see which linear function f is closest too, we find the ℓ_a which maximizes $\langle f, \ell_a \rangle$.

Definition 9 *$\widehat{f}_a$, the* ath Fourier coefficient of f *is defined to be $\langle f, \ell_a \rangle = 1 - 2\Pr[f(x) \neq \ell_a(x)]$.*

¿From Parseval's identity we get a bound on how large these correlations between f and linear functions ℓ_a can be:

Proposition 10 $\sum_{a \in \{0,1\}^m} \widehat{f}_a^{\,2} = 1$.

Now that we have some understanding of how we can use Fourier coefficients to study how close a function is to being linear, let's examine how we can use Fourier coefficients to get a handle on the linearity test. Recall our definition for the degree to which f passes the linearity test:

Definition 11 *$\delta(f) = \Pr_{x,y}[f(x+y) = f(x)f(y)]$ (The right side has $f(x)f(y)$ rather than $f(x) + f(y)$ because of our switch from $\{0,1\}$ to $\{+1,-1\}$.)*

We have the following proposition, similar to Proposition 6:

Proposition 12 $\mathrm{Ex}_{x,y}[f(x)f(y)f(x+y)] = 1 - 2\delta(f)$

Proof

The linearity test accepts x, y iff $f(x)f(y) = f(x+y)$ iff $f(x)f(y)f(x+y) = 1$, since f takes values in $\{+1, -1\}$. Thus:

$$\begin{aligned} \mathrm{Ex}_x[f(x)f(y)f(x+y)] &= (+1) \cdot \Pr_{x,y}[f(x)f(y)f(x+y) = 1] + (-1) \cdot \Pr_{x,y}[f(x)f(y)f(x+y) = -1] \\ &= 1 - 2\delta(f) \end{aligned}$$

■

Now we wish to show that if $\delta(f)$ is small for a given function then $\epsilon(f)$ is also small — i.e., f has at least one large Fourier coefficient. We do this by "projecting" f onto our basis functions $\{\ell_a\}$. Thus we should first decide how the expectation in the previous proposition works out when the f's are replaced with arbitrary linear functions.

Proposition 13 $\mathrm{Ex}_{x,y}[\ell_a(x)\ell_b(y)\ell_c(x+y)]$ *is 1 if* $a=b=c$*, and is 0 otherwise.*

Proof

In case that $a=b=c$, $\ell_a(x)\ell_b(y)\ell_c(x+y)$ is always 1 since ℓ_a is linear. Otherwise, we may assume without loss of generality that $a \neq b$. In this case, we have:

$$\begin{aligned} \mathrm{Ex}_{x,y}[\ell_a(x)\ell_b(y)\ell_c(x+y)] &= \mathrm{Ex}_{x,y}[\ell_a(x)\ell_b(y)\ell_c(x)\ell_c(y)] \\ &= \mathrm{Ex}_x[\ell_a(x)\ell_c(x)]\mathrm{Ex}_y[\ell_b(y)\ell_c(y)], \end{aligned}$$

the first equality holding because ℓ_c is linear.

Now note that if $a=0$ then $\ell_a(x)=1$ for all x so $\mathrm{Ex}_x[\ell_a(x)]=1$. However, if $a\neq 0$ then we have $a_i \neq 0$ for some i. Then we can pair up all of the strings in $\{0,1\}^m$, matching strings which differ only in the ith place. For such a pair of strings x, x', we will have $\ell_a(x)=+1$, $\ell_a(x)=-1$, or the other way around; regardless, $\ell_a(x)+\ell_a(x')$ will be 0, and it thus follows that $\mathrm{Ex}_x[\ell_a(x)]$ must be 0.

Consequently, $\mathrm{Ex}_x[\ell_a(x)\ell_c(x)]$ is just $\langle \ell_a, \ell_c\rangle$, and $\mathrm{Ex}_x[\ell_b(x)\ell_c(x)]$ is just $\langle \ell_b, \ell_c\rangle$. But since $a \neq b$, exactly one of these two inner products is 0; hence their product is 0. Thus $\mathrm{Ex}_{x,y}[\ell_a(x)\ell_b(y)\ell_c(x+y)]=0$, as claimed.

■

Now we can do the projection we want, and calculate the expectation $\mathrm{Ex}_{x,y}[f(x)f(y)f(x+y)]$ from Proposition 12 in terms of Fourier coefficients.

Proposition 14 $\mathrm{Ex}_{x,y}[f(x)f(y)f(x+y)] = \sum_a \widehat{f_a}^3$.

Proof

This is just a calculation:

$$\begin{aligned} \mathrm{Ex}_{x,y}[f(x)f(y)f(x+y)] &= \mathrm{Ex}_{x,y}[(\textstyle\sum_a \widehat{f}_a\ell_a(x))(\sum_b \widehat{f}_b\ell_b(x))(\sum_c \widehat{f}_c\ell_c(x+y)) \\ &= \textstyle\sum_{a,b,c} \widehat{f}_a\widehat{f}_b\widehat{f}_c \mathrm{Ex}_{x,y}[\ell_a(x)\ell_b(y)\ell_c(x+y)] \qquad \text{(by linearity)} \\ &= \textstyle\sum_a \widehat{f_a}^3 \qquad , \end{aligned}$$

since the expectation is 0 unless $a=b=c$, from the previous proposition. ■

Corollary 15 $1-2\delta(f) = \sum_a \widehat{f_a}^3$.

Now we're essentially done:

Theorem 16 *For all code functions f, $\epsilon(f) \leq \delta(f)$ — i.e., if f fails the linearity test with probability at most δ, then it is δ-close to a linear function.*

Proof

$$\begin{aligned}
\sum_a \widehat{f_a}^3 &\leq \sum_a (\max_b \widehat{f_b})\widehat{f_a}^2 \\
&= (\max_b \widehat{f_b}) \sum_a \widehat{f_a}^2 \\
&= \max_b \widehat{f_b},
\end{aligned}$$

since $\sum_a \widehat{f_a}^2 = \langle f, f \rangle = 1$, by Parseval's identity and the fact that f takes values in $\{+1, -1\}$.

Thus we have $1 - 2\delta(f) = \sum_a \widehat{f_a}^3 \leq \max_b \widehat{f_b} = 1 - 2\epsilon(f)$, and the theorem follows.

■

So this theorem gives us an excellent bound on how close a function is to being linear, if it passes the linearity test with high probability.

Compare now our correlation of $\epsilon(f)$ and $\delta(f)$ from the last lecture with the one we have now:

(In these two figures, the possible values for δ and ϵ are those that lie above the curve and the straight line.)

We have a much better comparison now. With reference to the lower figure, our theorem shows that the only feasible values for δ and ϵ are those that lie on or above the line of slope 1. One can actually do a little bit better as indicated in the figure, bounding δ above $45/128$. It's also possible to show that the highlighted point on the curve with $\delta = 45/128$ is achievable, so the bound is tight there.

What is the significance of the other highlighted point, $(\epsilon, \delta) = (\frac{1}{2}, \frac{1}{2})$? Note that two random functions with range $\{+1, -1\}$ agree with probability $\frac{1}{2}$. But if we get a function which passes the linearity test with probability just a little better than $\frac{1}{2}$, then we get a nontrivial correlation between the function and a linear function. Is a very small correlation really interesting though? Yes, in fact. Suppose f passes the linearity test with probability $\frac{1+\epsilon}{2}$. Then by our theorem, there exists an a such that $\Pr[f(x) \neq \ell_a(x)] \leq \frac{1-\epsilon}{2}$. By definition, this happens if and only if there exists an a such that $\widehat{f_a} \geq \epsilon$. But recall from Parseval's identity we had $\sum_{a \in \{0,1\}^m} \widehat{f_a}^2 = 1$; thus at most $\frac{1}{\epsilon^2}$ possible a's can satisfy $\widehat{f_a} \geq \epsilon$. Thus we really do have an interesting result since we can say that f isn't close to too many linear functions — at most $\frac{1}{\epsilon^2}$ of them. As it happens, this bound of $\frac{1}{\epsilon^2}$ is more or less asymptotically tight as $\epsilon \to 0$.

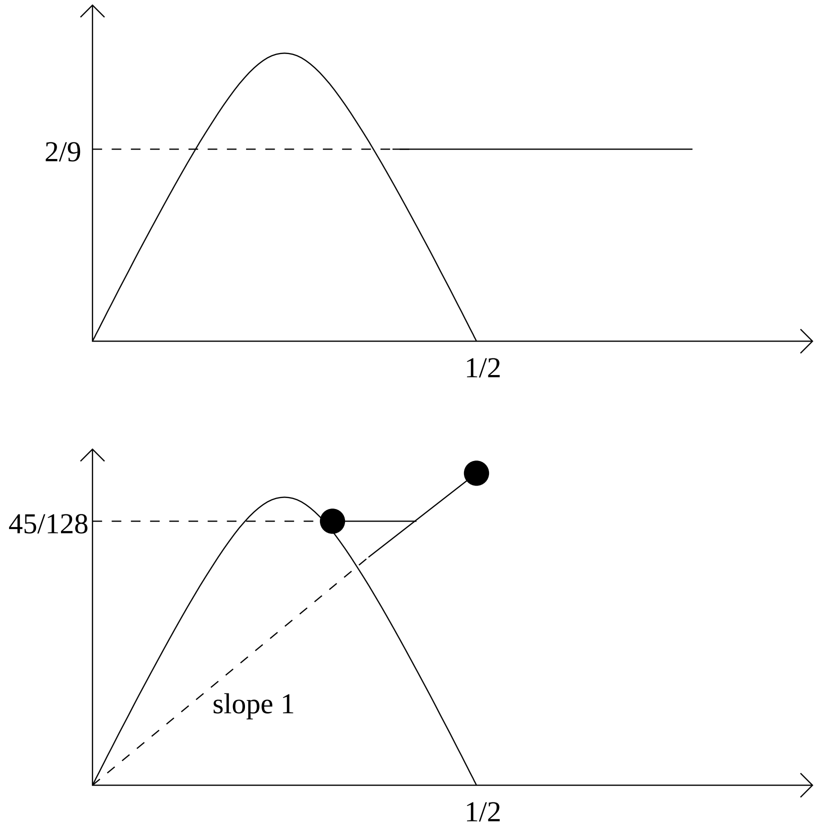

Figure 1: Our old and new correlations of $\delta(f)$ (y-axis) with $\epsilon(f)$ (x-axis).

2 One More View On Linearity Testing, Via Duality In Coding Theory

In this section we briefly sketch a third idea for linearity testing; it can be used in general over finite fields, but we restrict attention to GF(2). The idea, based on duality in coding theory, is due to Kiwi.

Given a function $f : \{0,1\}^m \to \mathrm{GF}(2)$, we define a code C_f. Its codewords are strings in $\{0,1\}^{2^m}$, and they are given by the set of linear functions ℓ_a on $\{0,1\}^m$, along with all functions of the form $f + \ell_a$. Certainly C_f contains the Hadamard code. Further, if f is a linear function itself, then $\{f + \ell_a : \ell_a \text{ linear}\}$ is just the set of linear functions, so C_f *is* the Hadamard code. But if f is merely very close to linear, then C_f contains two codewords which are very close. Indeed, it's easy to see:

Proposition 17 *The relative minimum distance of C_f is $\epsilon(f)$; i.e., the minimum distance of the code is $\epsilon(f)$ times the length of the code, 2^m.*

Now we use some coding theory to analyze the situation. Let $B \subseteq \{0,1\}^n$ be any linear code, i.e., one with the property that the sum of two codewords is a codeword.

Definition 18 *The* dual *of the code B is the code $B^{\perp} = \{c : \langle c, b\rangle = 0 \ \forall b \in B\}$. Here the inner* *product $\langle c, b\rangle$ is just $\sum c_i b_i$ mod 2.*

Definition 19 *For the code B, the* weight distribution *of B is the sequence $\langle \beta_0, \beta_1, \ldots, \beta_n\rangle$, where β_i is the number of codewords in B which have exactly i nonzero coordinates.*

A special case of the MacWilliams identity relates the weight distribution of a code B to that of its dual $B^\perp$, and vice versa. Now it turns out we have some partial information about the weight distribution of C_f; namely, it is $\langle 1, 0, 0, \ldots, 0, \geq 0, \geq 0, \ldots, \geq 0\rangle$, where the ≥ 0 coordinates do not begin until at least the $\epsilon(f)2^m$ coordinate. Why does this help? Kiwi's primary observation is that a dual codeword to a Hadamard code specifies a kind of linearity test. In particular, the linearity tests we've been using correspond to dual codewords of weight 3. I.e., f passes a certain one of our linearity tests if and only if the associated dual codeword is in $C_f^\perp$. So $\delta(f)$ is related to the number of codewords of weight 3 in $C_f^\perp$. But as we mentioned, the MacWilliams identity relates the weight distribution of $C_f^\perp$ with the weight distribution of C_f; the former weight distribution involves δ and the latter weight distribution involves ϵ. From this, Kiwi is able to relate ϵ and δ.