PCPs and HDX - Lecture 2

November 15, 2016

In the previous lecture we stated the PCP theorem and related it to a generalized notion of expansion. In the first part of the course we will be constructing PCPs and constraint expanders. Today's construction is known as the *linearity testing* construction. We will continue the "expansion" point of view, for a general system of constraints, and construct an explicit expanding system of constraints.

1 Expansion of a constraint system

Let V be a finite set. A constraint is specified by a tuple $(v_1, \ldots, v_q)$ and a predicate $\varphi : \Sigma^q \to \{0,1\}$. We think of the constraint as restricting the set of functions $f : V \to \Sigma$ by telling us which values on $v_1, \ldots, v_q$ are valid. Formally, a function $f : V \to \Sigma$ satisfies the constraint $((v_1, \ldots, v_q), \varphi)$ iff $\varphi(f(v_1), \ldots, f(v_q)) = 1$. The constraint is called *q-local* and we are interested in the case where $q \ll n$, so a constraint looks at very few values of the function.

In the previous lecture we studied two systems of constraints, one corresponding to 3-Coloring and the other corresponding to graph expansion. The constraints in those systems were always on pairs of variables, i.e. they were 2-local, which makes it easy to draw them on a graph with vertices and edges. Many well-studied systems of local constraints have higher locality. For example 3SAT is the a system of constraints that are 3-local over alphabet $\Sigma = \{0,1\}$, and each constraint over three variables allows seven of the eight possible assignments to the variables.

Today we will discuss another system with locality 3. Once we go above 2 the graph of constraints turns into a hypergraph, and the language of high dimensional graphs becomes relevant.

Definitions. Given a system $\mathbf{C}$ of constraints, we define $SAT(\mathbf{C})$ to be the set of functions $f : V \to \Sigma$ that satisfy every constraint in $\mathbf{C}$. For every $f : V \to \Sigma$ we can define the distance of f from $SAT(\mathbf{C})$ as

$$\mathrm{dist}(f, SAT(\mathbf{C})) = \min_{g \in SAT(\mathbf{C})} dist(f, g)$$

where $\mathrm{dist}(f, g)$ is the fraction of entries on which f, g differ. The fraction of constraints rejecting f is

$$\mathrm{rej}_{\mathbf{C}}(f) = \Pr_{c \in \mathbf{C}}[c \text{ rejects } f]$$

Clearly

$$f \in SAT(\mathbf{C}) \quad \Leftrightarrow \quad \mathrm{dist}(f, \mathrm{SAT}(\mathbf{C})) = 0 \quad \Leftrightarrow \quad \mathrm{rej(f)} = 0$$

When $f \notin SAT(\mathbf{C})$ then both $\mathrm{dist}(f, \mathrm{SAT}(\mathbf{C})) > 0$ and $\mathrm{rej}(f) > 0$. We are interested in the ratio between the two, minimized over all $f \notin SAT(\mathbf{C})$,

$$\gamma(\mathbf{C}) = \min_{f \notin SAT(\mathbf{C})} \frac{\mathrm{rej}_{\mathbf{C}}(f)}{\mathrm{dist}(f, SAT(\mathbf{C}))}.$$

We are interested in constructing and studying systems of constraints that are expanding. We saw that every expander graph can be viewed as an expanding system of equality constraints. However, the property it defines, SAT(G), consists of only the constant functions, so it is not very interesting. Today we will construct our first system of constraints that is non-trivial. This is a first step towards constructing a PCP.

2 Hadamard Code and BLR Linearity Testing

Let $V = \{0,1\}^n$, and $\Sigma = \{0,1\}$. We will consider all Boolean functions, $f : V \to \{0,1\}$, that satisfy *linearity constraints*.

Blum, Luby, and Rubinfeld (BLR) suggested the following set of linearity constraints. For each $x, y \in V$ we will have a constraint checking that

$$f(x) + f(y) = f(x + y)$$

Let $\mathbf{BLR}_n$ be the set of all constraints above, where n is the parameter for the dimension of the space $V = \{0,1\}^n$. It is easy to see that the set of functions that satisfy all of the constraints is

$$\mathrm{SAT}(\mathbf{BLR}_\mathrm{n}) = \{\mathrm{f} : \{0,1\}^\mathrm{n} \to \{0,1\} \mid \mathrm{f \text{ is a linear function}}\}$$

In other words,

Claim 2.1. *Let $f : \{0,1\}^n \to \{0,1\}$. If $\mathrm{rej}(f) = 0$ then f is linear, i.e. there is some $a \in \{0,1\}^n$ such that for all $x \in \{0,1\}^n$, $f(x) = \sum_i a_i x_i \mod 2$.*

The reason for the name *Hadamard code* is because the Hadamard encoding is the name for the encoding that sends a message $\vec{a} = (a_1, \ldots, a_n) \in \{0,1\}$ to the linear function whose coefficients are $a_1, \ldots, a_n$. It encodes an n bit string by an 2^n bit string that is the truth table of the linear function whose coefficients are $\vec{a}$. We will prove the following theorem,

Theorem 2.2. *For all n, $\gamma(\mathbf{BLR}_n) \geq \frac{2}{9}$. In other words, for every $f : \{0,1\}^n \to \{0,1\}$, $\mathrm{rej}_{\mathbf{BLR}_n}(f) \geq \frac{2}{9} \cdot \mathrm{dist}(f, \mathrm{SAT}(\mathbf{BLR}_\mathrm{n}))$.*

This theorem says that the BLR constraints are expanding. Usually it is viewed as saying that they provide a local test for whether a given function is a valid Hadamard encoding.

Proof. Since $\mathrm{dist}(f, \mathrm{SAT}(\mathbf{BLR}_\mathrm{n})) \leq 1$, it is enough to prove the theorem assuming that $\varepsilon = \mathrm{rej}(f) < 2/9$. We need to prove that if $\varepsilon = \mathrm{rej}(f)$ is small then f is close to a linear function. We will "correct" f by changing it into a function g and then argue that g is linear, and that it is close to f.

Majority decoding. For each $x \in \{0,1\}^n$ let $g(x)$ be the value for $f(x)$ that would cause more tests (that look at the point x to accept than to reject. I.e. define

$$g(x) = \mathrm{popular}_{y \in \{0,1\}^n}[f(y) + f(x + y)] \tag{1}$$

where $\mathrm{popular}_x[f(x)]$ denotes the most popular value of f ranging over all x. Define the probability of the popular vote for x as

$$P_x = \Pr_y[f(y) + f(x + y) = g(x)]$$

Claim 2.3. *If $\mathrm{rej}(f) < 2/9$ then for all $x \in \{0,1\}^n$, $P_x > 2/3$.*

Proof. Choose y and z uniformly at random. On one hand, the probability over y that $f(y)+f(x+y)=g(x)$ is P_x, and the probability that $f(z)+f(x+z)=g(x)$ is also P_x. The probability that $f(y)+f(x+y)=f(z)+f(x+z)$ is thus $(P_x)^2+(1-P_x)^2$ because y and z are independent. Switching terms around this holds iff

$$f(y)+f(z)=f(x+y)+f(x+z)$$

Now with probability at least $1-\varepsilon$, $f(y)+f(z)=f(y+z)$ and similarly with probability $1-\varepsilon$, $f(x+y)+f(x+z)=f(y+z)$. So by a union bound the probability that both hold (now there is no independence!) is at least $1-2\varepsilon \geq 5/9$. So

$$P_x^2+(1-P_x)^2 > 5/9$$

and this can only hold if $P_x < 1/2$ or $P_x > 2/3$. The first is ruled out since $P_x \geq 1/2$. □

Claim 2.4. *The function g defined in (1) satisfies all the constraints in $\mathbf{BLR}_n$, hence it is linear.*

Proof. Fix x, y. By the previous claim, $P_x > 2/3, P_y > 2/3$ and $P_{x+y} > 2/3$, so each of the following three equations hold with probability above $2/3$ over the choice of z:

$$g(x)=f(z)+f(x+z)$$

$$g(y)=f(z)+f(y+z)$$

$$g(x+y)=f(x+z)+f(y+z)$$

(for the third equation note that choosing a random z and setting $w=x+z$ gives a uniformly distributed w and then $f(w)+f(w+(x+y))=f(x+z)+f(y+z)$). So there must be some z_0 for which all three equations hold simultaneously. If we sum them up we get identically zero on the right hand side, implying that $g(x)+g(y)+g(x+y)=0$. □

Finally, for every x where $f \neq g$ we have that the test rejects on at least $2/3$ of the choices of y, so

$$\begin{aligned}
\mathrm{rej}(f) &= \mathbb{E}_x \Pr_y[f(x) \neq f(y)+f(x+y)] \\
&\geq \Pr_x[f(x) \neq g(x)] \cdot \mathbb{E}_{x:f(x)\neq g(x)} \Pr_y[f(x) \neq f(y)+f(x+y)] \\
&= \mathrm{dist}(f,g) \cdot \mathbb{E}_{x:f(x)\neq g(x)} P_x \\
&\geq \mathrm{dist}(f,g) \cdot \frac{2}{3}
\end{aligned}$$

where in the first inequality we used the formula $\mathbb{E}[A] \geq \Pr[B] \cdot \mathbb{E}[A|B]$. As long as $\mathrm{rej}(f) \leq 2/9$ we have shown that

$$\mathrm{rej}(f) \geq \frac{2}{3} \cdot \mathrm{dist}(f,g) \geq \frac{2}{3}\mathrm{dist}(f, \mathrm{SAT}(\mathbf{BLR}_\mathrm{n}))$$

and altogether $\mathrm{rej}(f) \geq \min(\frac{2}{9}, \mathrm{dist}(f, \mathrm{SAT}(\mathbf{BLR}_\mathrm{n})) \cdot \frac{2}{3}) \geq \frac{2}{9}\mathrm{dist}(\mathrm{f}, \mathrm{SAT}(\mathbf{BLR}_\mathrm{n}))$. □

3 More constraint-expanders ?

We are interested in constructing systems of constraints that are expanding. How easy is it to get such a system? is a random system a good choice? It is known that a random d-regular graph is with high probability an expander. It is natural to wonder if for certain parameters, a random system of constraints is a constraint-expander. The short answer is no: If we choose the system of constraints $\mathbf{C}$ at random, then if their number is a large enough linear multiple of n the property $\mathrm{SAT}(\mathbf{C})$ becomes empty. If we choose fewer constraints then the system turns out to be non expanding. (To see this, think of removing a single constraint, getting $\mathbf{C}^-$ from $\mathbf{C}$. If $x \in \mathrm{SAT}(\mathbf{C}^-) \setminus \mathrm{SAT}(\mathbf{C})$ then it violates only a single constraint in $\mathbf{C}$. It can be shown that x could still whp be quite far from the set $\mathrm{SAT}(\mathbf{C})$).

It is an open question to find a constraint expander $\mathbf{C}$ whose property $\mathrm{SAT}(\mathbf{C})$ has constant rate and distance. These terms come from coding theory. Rate refers to $\frac{\log |\mathrm{SAT}(\mathbf{C})|}{N}$ where N is the length of the strings in $\mathrm{SAT}(\mathbf{C})$ ($N = 2^n$ in the case of the Hadamard code). Distance refers to $\min_{f \neq g \in \mathrm{SAT}(\mathbf{C})} \mathrm{dist}(f, g)$. We have seen a first construction of a non-trivial constraint expander. The distance is a constant (it is $\frac{1}{2}$), but the rate is $\log 2^n / N = n/N = (\log N)/N$. In the next lectures we will see constructions with higher rates.