PCP - Lecture 8

* Linearity Testing
* $NP \subseteq PCP[poly, 1]$

Note Title 4/15/2008

Our next goal in the course is to prove

$NP \subseteq PCPP[poly, 1]$ ← Note the constant query + binary alphabet

proximity

We will be able to use this verifier in a composition scheme to prove the PCP thm.

First however, we show a verifier not for NP but for a simpler language - consisting of all linear functions in n variables over $\{0,1\}$

Linearity Testing

Given $f: \{0,1\}^n \to \{0,1\}$
We wish to test whether f is linear. Making few queries to f.

Def: $f: \{0,1\}^n \to \{0,1\}$ is linear if
$\forall x,y \in \{0,1\}^n \; f(x)+f(y) = f(x+y)$ (addition modulo 2)
(of course, this def extends to any group G replacing $\{0,1\}$.)

Claim: f is linear iff $\exists a \in \{0,1\}^n$ s.t. $f(x) = \langle a, x\rangle = \sum_i a_i x_i \bmod 2$.

Proof: ($\Leftarrow$) clear. For ($\Rightarrow$): let $a_i = f(e_i)$ for $e_i = (0 \ldots 0, \overset{i}{1}, 0 \ldots 0)$.
and the claim follows by linearity.

Clearly, # queries is at least 3. We now show a 3-query test:

BLR Test: Choose random $x, y \in \{0,1\}^n$.

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

Clearly if f linear $\Rightarrow$ $\Pr(\text{success}) = 1$.

What happens if f is not linear? Need to talk about distance.

Def: Let $f, g : \{0,1\}^n \to \{0,1\}$, denote $\text{dist}(f,g) = \Pr[f(x) \neq g(x)]$

f, g are δ-far if $\text{dist}(f,g) \geq \delta$, δ-close if $\text{dist}(f,g) \leq \delta$

f is δ-far from linear if it is δ-far from all linear g.

Theorem: If f is δ-far from linear,

$$\text{then } \Pr[T \text{ rejects } f] \geq \min\left(\frac{2}{9}, \frac{\delta}{2}\right) \geq \frac{2}{9}\delta$$

Comments:

* "special case" of "low degree test". Preceded it historically.
* can be extended to groups homomorphism testing.
  $\exists$ groups for which $\frac{2}{9}$ is tight!
* For our case, $\{0,1\}$, can prove, via Fourier analysis that
  $\Pr[T \text{ rejects}] \geq \delta$.

: let $f : \mathbb{Z}_9^n \to \mathbb{Z}_9$. $f(u) = 3k$ if $u_1 = 3k, 3k-1, 3k+1$

prove (a) $\Pr[T \text{ rej } f] = \frac{2}{9}$

(b) $\text{Dist}(f, \text{Lin}) = 2/3$.

Proof: Note $f(x)+f(y) \neq f(x+y)$ iff $x_1 = y_1 = 1 \pmod 3$ or $x_1 = y_1 = -1 \pmod 3$

this happens w. prob. $2/9$.

$\text{Dist}(f, \text{Lin}) = 2/3$; ...

Proof of Theorem: We use a correction to majority argument.

Idea: Define a "corrected" version of f, g:

For each x consider $f(y) + f(x+y)$ for all y.

Define $g(x) = 1$ if $\Pr_y\left[f(y) + f(x+y) = 1\right] \geq \frac{1}{2}$, and $g(x) = 0$ o.w.

Also let $P_x = \text{Prob}_y\left[f(y) + f(x+y) = g(x)\right]$. Clearly $\frac{1}{2} \leq P_x \leq 1$.

Claim 1: $\text{Prob}\left[T \text{ rejects } f\right] \geq \frac{1}{2} \cdot \text{dist}(g, f)$

Proof:

$$\text{Prob}\left[T\text{ rej}\right] = \underbrace{\Pr\left[g \neq f\right]}_{= \delta} \underbrace{\Pr\left[T\text{ rej} \mid g \neq f\right]}_{\geq \frac{1}{2}} + \Pr\left[g = f\right] \Pr\left[T\text{ rej} \mid g = f\right]$$

$$\geq \delta / 2$$

$\square$

Claim 2: If $\Pr(T\text{ rej}) < \frac{2}{9}$ then $\forall x$ $P_x \geq \frac{2}{3}$.

Proof:

$$\text{Prob}_{y,z}\left[f(y) + f(x+y) = f(z) + f(x+z)\right] = (P_x)^2 + (1 - P_x)^2$$

rearranging

$$= \text{Prob}\left[\underbrace{f(y) + f(z)}_{= f(y+z) \text{ w. prob} \geq \frac{7}{9}} = \underbrace{f(x+y) + f(x+z)}_{= f(y+z) \text{ w. prob} \geq 7/9}\right] > \frac{5}{9}$$

so $(P_x)^2 + (1 - P_x)^2 > \frac{5}{9} \implies P_x > \frac{2}{3}$ $\square$

Claim 3: g is linear.

Proof:

$$\underbrace{f(z)\quad f(x+z)}_{\text{for } > \frac{2}{3}\ z\text{'s}} \qquad \underbrace{f(z)\quad f(z+y)}_{\text{for } > \frac{2}{3}\ z\text{'s}} \qquad \underbrace{f(z+x)\quad f(z-y)}_{\text{for } > \frac{2}{3}\ z\text{'s}}$$

$$g(x) + g(y) \qquad\qquad g(x+y)$$

$\exists z^*$ for which all three hold. $\Rightarrow$

$$\underbrace{f(z^*)+f(x+z^*)}_{g(x)} + \underbrace{f(z^*)+f(y+z^*)}_{g(y)} = \underbrace{f(x+z^*)+f(z^*+y)}_{g(x+y)}$$

$$g(x) + g(y) = g(x+y) \qquad \square$$

Conclusion: $\text{Prob}[T \text{ rejects}]$ is either $\geq \frac{2}{9}$ or

$$\text{Prob}[T \text{ rej}] \geq \frac{\delta}{2} = \frac{\text{dist}(f,g)}{2} \underset{g \text{ linear}}{\geq} \frac{\text{dist}(f,\text{Lin})}{2}.$$

$\blacksquare$

Def: The Hadamard Code maps to each $a \in \{0,1\}^n$ the linear function $L_a : \{0,1\}^n \to \{0,1\}$ defined by $L_a(x) = \sum_i a_i x_i \pmod 2$.

It is an error correcting code $H : \{0,1\}^n \longrightarrow \{0,1\}^{2^n}$ such that

(a) $\forall\, a \neq b \quad \text{dist}(L_a, L_b) = \frac{1}{2} \cdot 2^n$. (relative distance $= \frac{1}{2}$)

(b) Its rate is logarithmic ($\log N$ bits are mapped to N bits)

(c) It is locally testable with 3 queries.

it is a Locally Testable Code.

Research Question: Are there LTC's with good rate (linear?)

Thm #2: $\text{Prob}(T \text{ rejects } f) \geq \text{dist}(f, \text{Lin})$

Proof #2: (based on Fourier Analysis)

[we switch notation $0 \to 1$ $\quad 1 \to -1$ $\quad a \to (-1)^a$]

So a linear function is now $f(x) \cdot f(y) = f(xy)$ pointwise

Fix a function $f: \{\pm 1\}^n \to \{\pm 1\} \subseteq \mathbb{R}$. The space of all functions $f \in \mathbb{R}^{2^n}$ is a vector space.

The standard basis is $\{e_w\}_{w \in \{\pm 1\}^n}$: $e_w(x) = \begin{cases} 1 & x = w \\ 0 & \text{otherwise} \end{cases}$

Another basis is the following

$\forall S \subseteq [n]$ let $\chi_S(x_1 \ldots x_n) = \prod_{i \in S} x_i$. Clearly $\chi_S : \{\pm 1\}^n \to \{\pm 1\}$.

Define an inner product $\langle f, g \rangle = 2^{-n} \sum_x f(x) g(x)$.

(a) $\langle \chi_S, \chi_S \rangle = 1$

(b) $S \neq T$ $\quad \langle \chi_S, \chi_T \rangle = 2^{-n} \sum_x \prod_{i \in S} x_i \prod_{i \in T} x_i$

$$= 2^{-n} \sum_x \prod_{S \Delta T} x_i = 0$$

by pairing x according to their val on some coor in $S \Delta T$.

So $\{\chi_S\}$ is a basis, and $\forall f$ $\quad f = \sum_S \underbrace{\langle f, \chi_S \rangle}_{\text{notation: } \hat{f}_S} \cdot \chi_S$

(c) χ_S is linear: $\chi_S(x) \cdot \chi_S(y) = \prod_{i \in S} x_i y_i = \chi_S(xy)$.

(d) for any $f:\{\pm1\}^n \to \{\pm1\}$

$$\hat{f}_S = \langle f, \chi_S \rangle = \Pr(f=\chi_S) - \Pr(f\neq\chi_S) = 1-2\,\mathrm{dist}(f,\chi_S).$$

(e) for any $f:\{\pm1\}^n \to \{\pm1\}$ $\quad \sum (\hat{f}_S)^2 = 1$

$$\langle f,f\rangle = 2^{-n}\sum_x (f(x))^2 = 1$$

$$\langle f,f\rangle = \langle \sum_S \hat{f}_S\chi_S, \sum_S \hat{f}_S\chi_S \rangle = \sum \hat{f}_S^{\,2}$$

Now we want to relate $s = \Pr_{x,y}\left(f(x)f(y)\neq f(xy)\right) = \Pr_{xy}\left(f(x)f(y)f(xy)\neq 1\right)$ to the F. coef. of f.

Let $e = \mathbb{E}_{xy}\left(f(x)f(y)f(xy)\right)$ then $e = s-(1-s) = 1-2s$.

$$e = \mathbb{E}_{xy}\left[\sum_S \hat{f}_S\chi_S(x) \sum_T \hat{f}_T\chi_T(y) \sum \hat{f}_U\chi_U(xy)\right]$$

$$= \mathbb{E}_{xy} \sum_{STU} \hat{f}_S\hat{f}_T\hat{f}_U \prod_{i\in S} x_i \prod_{i\in T} y_i \prod_{i\in U} x_iy_i$$

$$= \mathbb{E}_{xy} \sum_{STU} \hat{f}_S\hat{f}_T\hat{f}_U \prod_{i\in S\Delta U} x_i \prod_{i\in T\Delta U} y_i = \sum_S (\hat{f}_S)^3$$

$$\leq \max_S \hat{f}_S \cdot \overbrace{\sum \hat{f}_S^{\,2}}^{=1} = \max_S \hat{f}_S = \hat{f}_{S_0}$$ (denoting S_0 a maximal coef)

$$\Rightarrow s = \frac{1-e}{2} \geq \frac{1-\hat{f}_{S_0}}{2} = \frac{1-(1-2\,\mathrm{dist}(f,\chi_{S_0}))}{2} = \mathrm{dist}(f,\chi_{S_0})$$

$$= \mathrm{dist}(f,\mathrm{Lin})$$ ∎

This completes our anlysis of the BLR linearity testing.

Self - Correction

As in the low degree case, the Had code allows self correction.

Lemma: $f:\{0,1\}^n \to \{0,1\}$, $\text{dist}(f, \text{Lin}) \leq \delta < \frac{1}{4}$.

Then (a) There is a unique linear $g:\{0,1\}^n \to \{0,1\}$ that is closest to f.

(b) There is a randomized two-query procedure S that guarantees for every x: $\Pr_r[S(x)=g(x)] \geq 1-2\delta > \frac{1}{2}$.

Proof: (a) If there were $g_1\ g_2$ linear, both δ-close to f then (by Δ ineq.:)

$\frac{1}{2} \leq \text{dist}(g_1, g_2) \leq \text{dist}(g_1, f) + \text{dist}(f, g_2) \leq 2\delta < \frac{1}{2}$

contradiction.

(b) $S(x)$: choose random $y \in \{0,1\}^n$, output $f(y) + f(x+y)$.

Since y is unif. dist. it hits the set $BAD = \{x \mid g(x) \neq f(x)\}$ with Prob. $\leq \delta$, and similarly $x+y$. Altogether:

$$\text{Prob}(f(y)+f(x+y) = g(x)) \geq \text{Prob}(y \notin B \text{ OR } x+y \notin B) \geq 1-2\delta > \frac{1}{2}.$$

... good for program checking.

Def: An ecc $C:\{0,1\}^k \to \{0,1\}^n$ is locally decodable with q-queries if there is a randomized procedure D such that, given $w \in \{0,1\}$, $\text{dist}(w, C) \leq \delta$, on input i D outputs x_i where $C(x)$ is the codeword closest to w. s.t. D makes only $\leq q$ queries into w.

Claim: The Hadamard Code is locally decodable with 2 queries.

Rest: On input i D runs $S(e_i)$ where $e_i = (0 \cdots 0, \overset{i}{1}, 0 \cdots 0)$.

research question: are there locally decodable codes with good rate?
(polynomial?) with 2-queries ← no!
3-queries ???

this is also a "popular" lower bound question

In fact, the Hadamard code allows one to read "correctly" not only the value of x_i, but also the value of $\ell(x)$ for all linear functions ℓ. (by the "self correction" property).

We will now slightly strenthen this property to all $q(x)$ for all quadratic functions $q(x_1 \ldots x_n) = \sum a_{ij} x_i x_j + \sum \ell_i x_i + \ell_0$.

Def: The quadratic functions encoding $Q: \{0,1\}^n \to \{0,1\}^{n^2}$ maps $(a_2 \ldots a_n) \in \{0,1\}^{n-1}$ into $H(a' \otimes a')$ where $a' \otimes a' \in \{0,1\}^{n^2}$ is defined by $(a' \otimes a')_{ij} = a'_i \cdot a'_j$ and $a' \in \{0,1\}^n$ is the vector $(1, a_2 \ldots a_n)$.

- The distance of this code is $\geq \frac{1}{2}$.
- The rate is $\sqrt{\lg n}$ ...
- locally testable? locally decodable?

Claim 1: Q is locally decodable. Moreover, for any quadratic function $q(x_1 .. x_n)$ there is a two-query procedure that whp gives $q(x_1 .. x_n)$ if given oracle access to $f \in \{0,1\}^{n^2}$ s.t. $\mathrm{dist}(f, Q) \leq \delta < \frac{1}{4}$.

Proof: Since $\mathrm{Im}(Q)$ is a subset of $\mathrm{Im}(\mathrm{Had})$ the local decoding procedure for Had works for Q. Moreover, every quadratic function q on $\vec{a}$ can be expressed as a linear function on $b = a' \otimes a'$. So by the strong self-correction of Q we get the result. ∎

Claim 2: Q is locally Testable.