6.895 Randomness and Computation February 13, 2006

Lecture 2

Lecturer: Ronitt Rubinfeld *Scribe: Iolanthe Chronis*

Lecture Outline:

- Linearity Testing (Introduction)
- Observation about distributions
- Self-correcting programs
  - union bound
  - Chernoff bound
  - indicator variables
  - linearity of expectations
- Proposed test for linearity
  - sampling claim - review "gap" bound
  - Coppersmith's example
  - proof of correctness and review of Markov's inequality (we didn't get to this)

1 Linearity Testing – An Introduction

Let G be a finite abelian group (ie $\mathbb{Z}_p$). The theorems that we will prove today work for arbitrary finite groups, but our proofs may not.

Definition 1 *A function $f : G \to G$ is linear (homomorphic) if $\forall x, y \in G$, $f(x) + f(y) = f(x + y)$.*

Examples of linear functions:

$$f(x) = x \tag{1}$$

$$f(x) = 3x \ (\mathrm{mod}\ l) \tag{2}$$

Problem: Can we tell if an arbitrary function f is linear? The function f is a black box. We know nothing about its internal structure; all we can do is query it, meaning we can pass in a value of x and get out a value of $f(x)$. To find out whether f is linear, we need to query every single value in the domain. If we do not, the following situation can occur:

$$f(x) = x \text{ for } x \neq 3, f(3) = 2 \tag{3}$$

There's no way to know that f isn't linear unless we query $x = 3$ itself.

Definition 2 *A function f is called ϵ-linear if there exists a function g s.t. g is linear and the following equation is true:*

$$\frac{\textit{number of } x\textit{'s s.t.} f(x) = g(x)}{\textit{number of } x\textit{'s}} = Pr_{x \in G}[f(x) = g(x)] \geq 1 - \epsilon \tag{4}$$

In the following, we will discuss how to test that f is ϵ-linear without testing all of the values in the domain.

2 An Observation About Distributions

If G is a finite group,

$$\forall a, y, \in G \ \Pr_x[y = a + x] = 1/|G| \tag{5}$$

Remark If we pick $x \epsilon_R G$ then $a + x \epsilon_R G$. We use this notation to denote that $a + x$ is distributed uniformly in G.

Since G is a group, inverses exist, so $y - a$ is a member of G, and only $x = y - a$ satisfies the equation $y = a + x$.

Remark If G is Z_2^n, addition is defined in the following manner:

$$(a_1, ..., a_n) + (b_1, ..., b_n) = (a_1 + b_1, ..., a_n + b_n) \tag{6}$$

3 Self-Correcting Programs

Say that f is the function described above that is linear except at $x = 3$. We can define a linear function g such that $g(i) = f(i)$ for all $i \neq 3$, and $g(3) = f(1) + f(2)$, or $g(3) = f(4) - f(1)$, or in general, $g(3) = g(x) + g(3 - x)$ for any $x \in G$ except for $x = 0$ or $x = 3$. This allows us to correct 1 error, but we can extend this to correct more as follows.

For a general ϵ-linear f where $\epsilon \leq 1/8$, let g be the closest linear function to f (the following shows that when $\epsilon \leq 1/8$, g is unique). If we pick y randomly from G, what is the probability that $g(x) = f(y) + f(x - y)$?

$$\text{Let cond (1) be } f(y) \neq g(y) \text{ (and } Pr_y[\text{cond (1)}] \leq \epsilon) \tag{7}$$
$$\text{Let cond (2) be } f(x - y) \neq g(x - y) \text{ (and } Pr_{x,y}[\text{cond (2)}] \leq \epsilon) \tag{8}$$
$$\Pr_{x,y}[\text{cond (1) or cond(2)}] \leq 2\epsilon \text{ by the union bound} \tag{9}$$
$$\text{We have that } \forall x, y, g(x) = g(y) + g(x - y) \tag{10}$$
$$\text{Therefore } \Pr_{x,y}[g(x) = f(y) + f(x - y)] \geq 1 - 2\epsilon \geq 3/4 \tag{11}$$

Using this bound, we can use the following algorithm to find the appropriate values for $g(x)$ using other parts of f.

Self-Corrector(x):
For $i = 1...r$
  Pick y randomly from G
  $answer_i \leftarrow f(y) + f(x - y)$
$output$ = most common $answer_i$

Theorem 3 *$Pr[output = g(x)] \geq 1 - \beta$ (for proper choice of constants.)*

Proof Let's define a random indicator variable $\sigma_i = 1$ if $answer_i = g(x)$, and 0 otherwise. Thus, the expected value of σ_1 is just the probability that $answer_i = g(x)$. In other words,

$$
\begin{aligned}
&\mathrm{E}[\sigma_i] = \Pr[\sigma_i = 1] \geq 3/4 && (12)\\
&\text{If } \Sigma\sigma_i > r/2 \text{ then } \mathit{output} = g(x) && (13)\\
&\mathrm{E}[\Sigma\sigma_i] = \Sigma E[\sigma_i] = 3r/4 && (14)\\
&\Pr[\sigma_i \leq r/2] \leq \Pr[|\frac{\Sigma\sigma_i}{r} - \frac{\mathrm{E}[\Sigma\sigma_i]}{r}| \leq 1/4] && (15)
\end{aligned}
$$

Now we can use a Chernoff bound. The mathematical statement of a Chernoff boundfor an indicator variable σ_i with $\Pr[\sigma_i]= p$ is[1]:

$$\Pr[|\frac{\Sigma\sigma_i}{r} - p| \leq \epsilon p] \leq 2e^{-rp\epsilon^2/3} \tag{16}$$

Here, $p = 3/4$ and $\epsilon p = 1/4$, so $\epsilon = 1/3$. Also, choose $r = c\ln(1/\beta)$. Plug these values into the Chernoff bound formula:

$$
\begin{aligned}
&2e^{-r(3/4)(1/9)/3} && (17)\\
&= 2e^{-r/36} && (18)\\
&= 2e^{-c/36\ln(1/\beta)} && (19)\\
&< \beta \text{ for } c < 36. && (20)
\end{aligned}
$$

■

4 A Proposed Test of Linearity

Here's a proposed test of linearity:
Repeat r times
  Pick $x, y \in_R G$
  If $f(x) + f(y) \neq f(x + y)$ output "fail"
Output "pass"

However, this example due to Coppersmith shows that a function can pass for most choices of x and y, but still be far from linear.

$$f(x) = \begin{cases} 1 & \text{if } x \equiv 1 \pmod 3 \\ 0 & \text{if } x \equiv 0 \pmod 3 \\ -1 & \text{if } x \equiv 2 \pmod 3 \end{cases}$$

This function fails for $x \equiv y \equiv 1 \pmod 3$ and for $x \equiv y \equiv 2 \pmod 3$, but it passes for all other cases. Thus, it fails for only $2/9$ of the possible choices of x and y. However, the closest linear function to f is the 0 function. The function f is only 0 in $1/3$ of all cases, so f is $2/3$-linear. It turns out that $2/9$ is a type of threshold in that functions that pass the linearity test *more* than $7/9$ of the time also are ϵ-linear for a fairly small ϵ. We will begin a proof of this idea at the end of this lecture, and finish the proof in the next lecture.

---

[1] See notes for better bounds

4.1 Sampling Theorem

Theorem 4 *There exists an algorithm s.t.*

if f is linear, i.e. $f(x) + f(y) = f(x+y) \forall x, y$
output "Pass"
if f is s.t. Pr[$f(x) + f(y) \neq f(x+y)$*]*$> \delta$
output "Fail" with probability $\geq 1 - \beta$
The runtime of this algorithm will be $O(\frac{1}{\delta} \ln \frac{1}{\beta})$

Bounds of this type will be refered to in this course as "gap" sampling bounds. This is not a term sanctioned by the greater community, since we will do many sampling tasks of this type later. "Gap" bound proofs always work in a manner similar to the following, which is the proof of the above theorem.

Remark $(1-x)^{1/x} \sim e^{-1}$

Proof Pr[output "Pass"] $\leq (1-\delta)^r = 1 - \delta^{\frac{c}{\delta} \ln \frac{1}{\beta}} = e^{-c \ln 1/\beta} = \beta^{-c} < \beta$ ■

Now we have shown that we can distinguish f's that are linear from those for which the test $f(x) + f(y) = f(x+y)$ often fails (with probability higher than δ. However, Coppersmith's example shows us that some f's that often pass that test (in the next lecture called "group law failure") are also very far from linear. However, for low enough δ, we will prove that f is close to linear.

Theorem 5 *Assuming Pr[*$f(x) + f(y) \neq f(x+y)$*]* $< \delta$ *and that G is a finite group,* $\delta < 2/9$ *implies that* f *is* $\delta/2$*-linear.*

Actually, we will prove this weaker version:

Theorem 6 *Assuming Pr[*$f(x) + f(y) \neq f(x+y)$*]* $< \delta$ *and that G is a finite abelian group,* $\delta < 1/16$ *implies that* f *is* 2δ*-linear.*

Definition 7 *If* $g(x) = plurality_{y \in G}[f(x+y) - f(y)]$*. We'll refer to* $f(x+y) - f(y)$ *as y's "vote".*

Definition 8 x *is called* ρ*-good if* $Pr_y[g(x) = f(x+y) - f(y)] > 1 - \rho$*. Otherwise,* x *is called* ρ*-bad.*

Note that if x is ρ-good for $\rho < 1/2$ then there is a "majority agreement."

Claim 9 *If* $\rho \leq 1/2$*,* $Pr_x[x$ *is* ρ*-good and* $g(x) = f(x)] > 1 - \delta/\rho$

We will prove the claim, and the rest of the theorem, in the next lecture.