**6.895 Randomness and Computation** February 11, 2008

# Lecture 2

*Lecturer: Ronitt Rubinfeld* *Scribe: Shira Mitchell*

## 1 Probabilistic Method

Last time we used randomness to speed up algorithms. Today we will use randomness to prove theorems by Erdös *probabilistic method*. The motto: if you probably exist, then you must exist. In other words, if the probability that an object exists is greater than zero, then it exists.

### 1.1 Existence of a 2-coloring

We begin by a simple example, demonstrating how the probabilistic method can be applied to the problem of showing the existence of a 2-coloring in sets.

Given subsets $S_1, \dots S_m \subseteq S$, $\forall i |S_i| = l$, is it possible to color each element of $S$ such that none of the $S_i$ are monochromatic (e.g. all red or all blue)? In general, you can't know without explicitly checking all of the colorings. However, we can show that the answer is always "yes" when $m < 2^{l-1}$.

**Theorem 1.** *If $m < 2^{l-1}$, there always exists a proper 2-coloring.*

*Proof.* Randomly assign a color (blue or red) to each element of $S$ independently with probability $\frac{1}{2}$ red and $\frac{1}{2}$ blue.

$$\begin{aligned}\forall i, \Pr[S_i \text{ is monochromatic}] &= \Pr[S_i \text{ is all blue}] + \Pr[S_i \text{ is all red}] \\ &= \frac{1}{2^l} + \frac{1}{2^l} = \frac{1}{2^{l-1}}.\end{aligned}$$

$$\begin{aligned}\Pr[\text{any } S_i \text{ is monochromatic}] &\leq \sum_i \Pr[S_i \text{ is monochromatic}] \\ &= m\frac{1}{2^l} < 1.\end{aligned}$$

Here we made use of the union bound, which says that $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$ for any two events $A$ and $B$. Since the probability that any of the $S_i$ is monochromatic is less than 1, the probability that none of the $S_i$ are monochromatic is greater than 0, showing the existence of such a coloring. □

### 1.2 Sum-free sets of integers

Let $A$ be a subset of positive integers (strictly $> 0$).

**Definition 2.** $A$ is *sum-free* if there do not exist $a_1, a_2, a_3 \in A$ such that $a_1 + a_2 = a_3$.

**Theorem 3** (Erdös '65). $\forall B = \{b_1, ..., b_n\}, \exists$ *sum-free* $A \subseteq B$ *such that* $|A| > n/3$.

For example, if $B = \{1, ..., n\}$, we can take $A = \{\lceil n/2 \rceil, ..., n\}$.

*Proof.* Without loss of generality, assume $b_n$ is the maximum, i.e. $b_n \geq b_i \; \forall i$.

Pick a prime $p \geq 2b_n$ such that $p \equiv 2 \pmod 3$, i.e. $p = 3k + 2$ for some integer $k$.
Let $C = \{k+1, ..., 2k+1\}$. Notice that

- $C$ is sum free,
- $C \subseteq \mathbb{Z}_p$,

- $|C|/(p-1) = (k+1)/(p-1) = (k+1)/(3k+1) > 1/3$.

Pick $x \in_R \{1, ..., p-1\} = \mathbb{Z}_p^*$. Set $d_i \equiv xb_i \pmod{p}$.

**Claim 4.** *$\forall i$ and for each $y \in \mathbb{Z}_p^*$, there is exactly one choice of $x \in \mathbb{Z}_p^*$ such that $y \equiv xb_i \pmod{p}$.*

*Proof.* First note that there exists such an $x$ since we can let $x \equiv yb_i^{-1}$ because $b_i$ is strictly positive and less than $p$, it is nonzero mod $p$, so it has an inverse mod $p$ ($\mathbb{Z}_p$ is a field). Since $y$ and $b_i$ are nonzero mod $p$, so is $x$. This $x$ is unique mod $p$, because if $x_1 b_i \equiv x_2 b_i \pmod{p}$, then multiplying by $b_i^{-1}$ gives $x_1 \equiv x_2 \pmod{p}$. □

Let $A_x$ be the preimage of $C$ under $x$. We know that $A_x$ is sum-free, because $b_i + b_j = b_k$ implies $xb_i + xb_j \equiv xb_k \pmod{p}$, and we get a contradiction since $C$ is sum-free.

How big is $A_x$? By the above claim, there are exactly $|C|$ choices of $x$ such that $xb_i$ is in $C$. Let $\sigma_i$ be 1, when $xb_i$ is in $C$, and 0, otherwise. Thus, $\Pr[\sigma_i = 1] = \mathrm{E}[\sigma_i] = |C|/(p-1) > 1/3$. By linearity of expectation, $\mathrm{E}[|A_x|] = \mathrm{E}[\sum_{i=1}^n \sigma_i] > n/3$. Thus, there exists an $x$ such that $|A_x| > n/3$ (since there has to be some value that is at least the average). □

How tight is this bound? 12/29 is know to be an upper bound. The exact optimum, which lies between 12/29 and 1/3, remains unknown.

# 2 The Boolean function $f : \{0,1\}^n \to \{0,1\}$

The Boolean function can be interpreted in a variety of ways, e.g.:

- as a truth table of a function (Complexity Theory),
- as a subset of the discrete cube (Coding Theory, Combinatorics),
- as a concept (Learning Theory).

We will consider all of these views in this course. We will develop tools from Fourier analysis for studying Boolean functions.

## 2.1 Linearity Testing

We begin with a specific task. Assume we have some black box that takes inputs $x$ and gives outputs $f(x)$. We can't look into the box and see anything about $f$. We know nothing about $f$'s internal structure; all we can do is query it, meaning we can pass in a value of $x$ and get out a value of $f(x)$.

**Definition 5.** *A function $f : \{0,1\}^n \to \{0,1\}$ is* linear *(or* homomorphic*) if $\forall x, y$, $f(x)+f(y) = f(x+y)$. (Where the first $+$ is over $\mathbb{Z}_2$ and the second $+$ is over $\mathbb{Z}_2^n$: $(a_1, ..., a_n)+(b_1, ..., b_n) = (a_1+b_1, ..., a_n+b_n)$.)*

Examples of linear functions:

- $f(x) = 0$
- $f(x) = x_1$
- $f(x) = (x_1 + x_{10}) \mod 2$
- $f_y(x) = x \cdot y = (\sum_{i=1}^n x_i y_i) \mod 2$, there are $2^n$ such functions, they are called *parity functions* and in fact, they are the only linear functions, because once you define $f$ on a basis, everything else is determined.

Problem: Can we tell if an arbitrary function $f$ is linear, if we view $f$ as a black box? To find out whether $f$ is linear, we need to query every single value in the domain. To see this suppose every query returns 0. If there is an input on which we don't evaluate the function, it may be the only input for which $f$ takes on 1, and our queries don't let us distinguish between linear and non-linear functions.

**Definition 6.** *A function $f : \{0,1\}^n \to \{0,1\}$ is* $\epsilon$-close to linear *if there exists a function $g$ s.t.*

- *$g$ is linear,*
- $\Pr[f(x) \neq g(x)] \leq \epsilon$.

*Otherwise we say that $f$ is* $\epsilon$-far from linear*.*

How can we test if $f$ is $\epsilon$-close to linear without testing all of the values in the domain ($2^n$ queries)?

### 2.1.1 Proposal

Here's a proposed test of linearity:

Repeat r $= \mathcal{O}(1/\rho \cdot \ln 1/\delta)$ times:
  Pick $x, y \in_R \{0,1\}^n$.
  If $f(x) + f(y) \neq f(x+y)$ output **FAIL**, and halt.
Output **PASS**.

### 2.1.2 Standard Claim

**Claim 7.**

1. *If f is linear, then the above algorithm always outputs* **PASS***.*
2. *If f is s.t. Pr[$f(x) + f(y) \neq f(x+y)$]$> \rho$, then $Pr[$output* **FAIL**$] \geq 1 - \delta$.

Part 1. of the above claim is obviously true, so it remains to prove Part 2.

*Proof of 2.* We will use the fact that $(1-x)^{1/x} \leq e^{-1}$.
$\Pr[\text{output "Pass"}] \leq (1-\rho)^r = (1-\rho)^{\mathcal{O}(1/\rho \ln 1/\delta)} = (1-\rho)^{c(1/\rho \ln 1/\delta)} \leq e^{-c \ln 1/\delta} = \delta^c \leq \delta.$ ☐

We have a characterization of linear functions: $\forall x, y\ f(x) + f(y) = f(x+y)$. We know that test we have presented rejects functions for which the above condition does not hold for many pairs $(x, y)$. How do we know that there are no linear functions that are $\epsilon$-far from any linear functions, and yet $f(x) + f(y) = f(x+y)$ does not hold for only very few pairs. We want a characterization of $\epsilon$-close to linear functions, and we would like to use $\Pr[f(x) + f(y) = f(x+y)]$ in it.

### 2.1.3 Great Notational Switch

We now consider $f : \{-1,1\}^n \to \{-1,1\}$ where we consider $-1$ to be like 1 and 1 to be like 0. We replace

| + | 0 | 1 |
|---|---|---|
| 0 | 0 | 1 |
| 1 | 1 | 0 |

with

| × | 1 | -1 |
|---|---|---|
| 1 | 1 | -1 |
| -1 | -1 | 1 |

**Definition 8.** *For $S \subseteq \{1, ..., n\}$, we define $\chi_S(x) = \prod_{i \in S} x_i$.*

Notice that if $x \in \{-1, 1\}^n$ has an even number of $-1$'s in $S$, then $\chi_S(x) = 1$, like in the old notation if $x \in \{0, 1\}^n$ had an even number of 1's in $S$ then $\sum_{i \in S} x_i \mod 2 = 0$. Similarly, if $x \in \{-1, 1\}^n$ has an odd number of $-1$'s in $S$, then $\chi_S(x) = -1$, like in the old notation if $x \in \{0, 1\}^n$ had an odd number of 1's in $S$ then $\sum_{i \in S} x_i \mod 2 = 1$.

### 2.1.4 Linearity Test

We now have

$$f(x)f(y)f(x \cdot y) = \begin{cases} 1 & \text{if test passes} \\ -1 & \text{if test fails} \end{cases}$$

This suggests the introduction of an indicator function

$$\delta(x, y) = \frac{1 - f(x)f(y)f(x \cdot y)}{2} = \begin{cases} 0 & \text{if test passes} \\ 1 & \text{if test fails} \end{cases}$$

which gives us a nice relation with the probability that a single execution of the test fails:

$$\Pr[\text{a single iteration of the test fails}] = E_{x,y}[\delta(x, y)].$$