Streaming Algs: how to handle massive data in small space

Property Testing: handle massive data in sublinear time.

Example Testing linearity

$f: \{0,1\}^n \to \{0,1\}$ is linear if $\boxed{f(x) = \sum_{i \in S} x_i \pmod 2}$ for some $S \subseteq [n]$

f is ε-far from linear if need to change $\varepsilon \cdot 2^n$ entries in truth table to get linear function

usually people treat ε as a constant

Input: truth table to f

Goal: decide if f is linear or far from linear

BLR Linearity Test [Blum Luby Rubinfeld '93]

Pick $x, y \in \{0,1\}^n$ at random

Query $f(x)$, $f(y)$, $f(x \oplus y)$

Reject if $f(x) + f(y) \neq f(x \oplus y)$

Note: if f linear

$$f(x) + f(y) = \sum_{i \in S} x_i + \sum_{i \in S} y_i$$

$$= \sum_{i \in S} x_i \oplus y_i$$

$$= f(x \oplus y)$$

Fact: if f linear BLR always accepts

if f ε-far from linear BLR rejects w/prob $\geq \varepsilon$. ← proof beyond scope of class. requires Fourier Analysis.

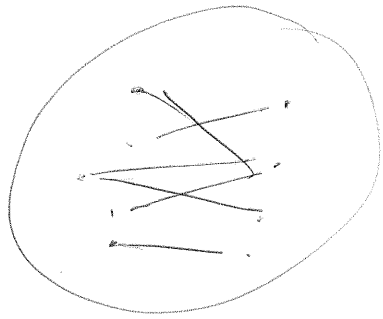
Solution: repeat BLR $O(\frac{1}{\varepsilon})$ times.

Accept if all tests pass.

Example testing bipartiteness

Input: undirected graph $G=(V,E)$

Goal: Is G bipartite or ε-far from bipartite?



↳ need to delete εN^2 edges to get bipartite graph

Bipartite ~~Property~~ Tester:

1. Pick random $S \subseteq V$ $|S| = k$
2. Query subgraph G_S: all (u,v): $u,v \in S$
3. Accept if G_S is bipartite
   Reject otherwise.

Theorem

- Bipartite-Test accepts all bipartite graphs
- if $k = \tilde{O}(1/\varepsilon^4)$ Bipartite-Test rejects ~~graphs~~ ε-far graphs w/prob $2/3$.

Property Testing Formal Model

Given a domain D (D = functions, graphs, ...)
A property is a subset $P \subseteq D$.

Defn Given metric $dist(a,b)$ on D, $\boxed{dist(a,P) := \min_{b \in P} dist(a,b)}$

a is ε-far from P if $\boxed{dist(a,P) \geq \varepsilon}$

Defn T is a property tester for P if

(1) if $a \in P$ then T accepts w/prob $2/3$ (completeness)
(2) if a ε-far from P then T rejects w/prob $2/3$ (soundness)

$Q'(P), Q^{NA}(P)$: min # queries needed for testers w/one-sided error, nonadaptive testers

$\boxed{\text{Query Complexity } Q(P) = \min \#\text{ queries needed to test } P}$

(3) T has one-sided error if it accepts all $a \in P$ (i.e. w/prob 1)
(4) T is nonadaptive if all queries chosen in advance

Lower Bounds in Property Testing

How do you prove lower bounds on query complexity?

The old way: Yao's Lemma + PAIN

The new way: Reduce From Communication Complexity

[Blais, Brody, Matulef '11]

Example: k-linearity

f is k-linear if $f(x) = \sum_{i \in S} x_i \pmod 2$ for some $S \subseteq [n]$, $|S| = k$

old bounds: $O(k \log k)$

$\Omega(\sqrt{k})$, $\Omega(k)$ for nonadaptive testers

new bound: $\Omega(k)$

proof: reduce from k-DISJ: Alice and Bob have sets $A, B \subseteq [n]$, $|A| = |B| = k$

want to check if $A \cap B \stackrel{?}{=} \emptyset$

Fact: $R(k\text{-DISJ}) = \Theta(k)$ [Hastad-Wigderson '07]

Let T be a tester for k-linearity.

k-DISJ protocol:

① Alice takes A, creates $f_A : \{0,1\}^n \to \{0,1\}$ $\quad f_A(x) = \sum_{i \in A} x_i$

Again, all addition is in $\mathbb{F}_2$ i.e., modulo 2

② Bob takes B, creates $g_B(x) := \sum_{i \in B} x_i$

③ Players use T to test

$$h := f_A \oplus g_B$$

for $(2k)$-linearity

$A = \{1, 3, 7\}$ $\quad f_A(x) = x_1 \oplus x_3 \oplus x_7$

$B = \{3, 5, 6\}$ $\quad g_B(x) = x_3 \oplus x_5 \oplus x_6$

$h(x) = (x_1 + x_3 + x_7) + (x_3 + x_5 + x_6)$

$= x_1 + x_5 + x_6 + x_7$

Emulating T:

Generate query x using public randomness

Alice → Bob $f(x)$

Bob sends $g(x)$

⇒ Both players compute $h(x)$

use public randomness to generate next query

Fact (1) if A, B disjoint then h is $(2k)$-linear

(2) if A, B intersect then h is $\not\equiv (k')$-linear for some $k' < 2k$

Claim Let $k' \neq k$. Then k'-linear functions are $\frac{1}{2}$-far from k-linear functions

proof Let $S, S' \subseteq [n]$ $|S| = k$, $|S'| = k'$ be given

$S \neq S'$ so $\exists\ i \in S \setminus S'$ (or vice versa)

For any $x \in \{0,1\}^n$, let $x^{(i)}$ be x w/ ith bit flipped

$x = 011001$

$x^{(4)} = 011101$

$i \in S$ so $f_S(x) \neq f_S(x^{(i)})$

$i \notin S'$ so $f_{S'}(x) = f_{S'}(x^{(i)})$

$\therefore$ either $f_S(x) \neq f_{S'}(x)$ or $f_S(x^{(i)}) \neq f_{S'}(x^{(i)})$

$\Rightarrow$ $f_S, f_{S'}$ differ on exactly half inputs. //

What have we proved?

(1) we can use property testing alg for k-linearity to create a communication protocol for k-DISJ

(2) the cost of the communication protocol is $2 *$ (#queries in tester)

(3) But $R(k\text{-DISJ}) = \Omega(k)$. $\Rightarrow$ $Q(k\text{-LIN}) \geq \frac{1}{2} R(k\text{-DISJ}) = \Omega(k)$ //

Testing Bipartiteness possible w/ $\tilde{O}(1/\varepsilon^4)$ queries

Algorithm

1. Pick random $S \subseteq V$ $|S| = \tilde{O}(1/\varepsilon^2)$
2. query (u,v) for all $u,v \in S$
3. ACCEPT if there is valid partition
   REJECT if (S_1, S_2) not bipartite $\forall$ partitions (S_1, S_2)

assume G is ε-far from bipartite

Proof we'll actually analyze this alg:

1. Pick ~~$\tilde{O}(1/\varepsilon)$~~ vertices $U \subseteq V$
2. Pick $S = \tilde{O}(|U|/\varepsilon)$ edges

query. How to partition $S = S_1 \cup S_2$?

Defn v is influential if it has $\frac{\varepsilon N}{4}$ neighbors

Claim w/prob $\geq 5/6$, ~~at least~~ all but at most $\frac{\varepsilon N}{4}$ influential ~~vertices~~ vertices adjacent to U

proof let v be influential

$\Pr[v \text{ not adj to } U] \leq (1 - \varepsilon/4)^{|U|}$

set $|U| = \frac{4 \ln 24/\varepsilon}{\varepsilon}$

$\leq e^{-\frac{|U|\varepsilon}{4}}$

$= e^{-\ln 24/\varepsilon}$

$= \frac{\varepsilon}{24}$

$\mathbb{E}[\#v \text{ not adj to } U] \leq \frac{N\varepsilon}{24}$

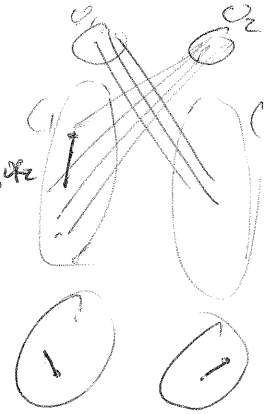
$\Pr[\geq \frac{N\varepsilon}{4} \; v \text{ not adj to } U] \leq \frac{1}{6}$ (Markov)

Now, Fix partition U_1, U_2 of U

C_1: neighbors of U_2

C_2: neighbors of U_1

R_1, R_2: arb. partition of remaining vertices



- Claim: $\leq \frac{\varepsilon N^2}{2}$ violating edges touch $R = R_1 \cup R_2$

proof $\leq \frac{\varepsilon N}{4}$ influential $\times N$

$+ \leq N$ non-influential $\times \frac{\varepsilon N}{4}$

$= \frac{\varepsilon N^2}{2}$ total edges

Claim: For any fixed partition $U_1 \cup U_2$, $\Pr[\exists$ partition S_1, S_2 that fails to witness non-bipartiteness$] \leq \frac{2^{-|U|}}{6}$

proof There are $\geq \frac{\varepsilon N^2}{2}$ violating edges in C.

$|S| = \frac{16|U|}{\varepsilon}$

We query ≥ 1 w/ prob $\geq 1 - (1 - \varepsilon/2)^{|S|/2} \geq 1 - \frac{2^{-|U|}}{6}$

if we hit one violating pair (v,w) how to partition it? w/o.g. say $v, w \in C_2$

can't put ~~both~~ $v, w \in S_1$, or $v, w \in S_2$

~~if one is put in S_1, then it violates~~

if we put $v \in S_1$ then (v,v) witnesses where $v \in U$ is v's neighbor in U_1.

Corollary Prob that we find ~~any~~ partition that doesn't witness non-bipartiteness $\leq 2^{|U|} \cdot \frac{2^{-|U|}}{6}$

$= 1/6$

Overall success: $2/3$ //