Nearly Optimal Sparse Group Testing

Venkata Gandikota, *Purdue University*

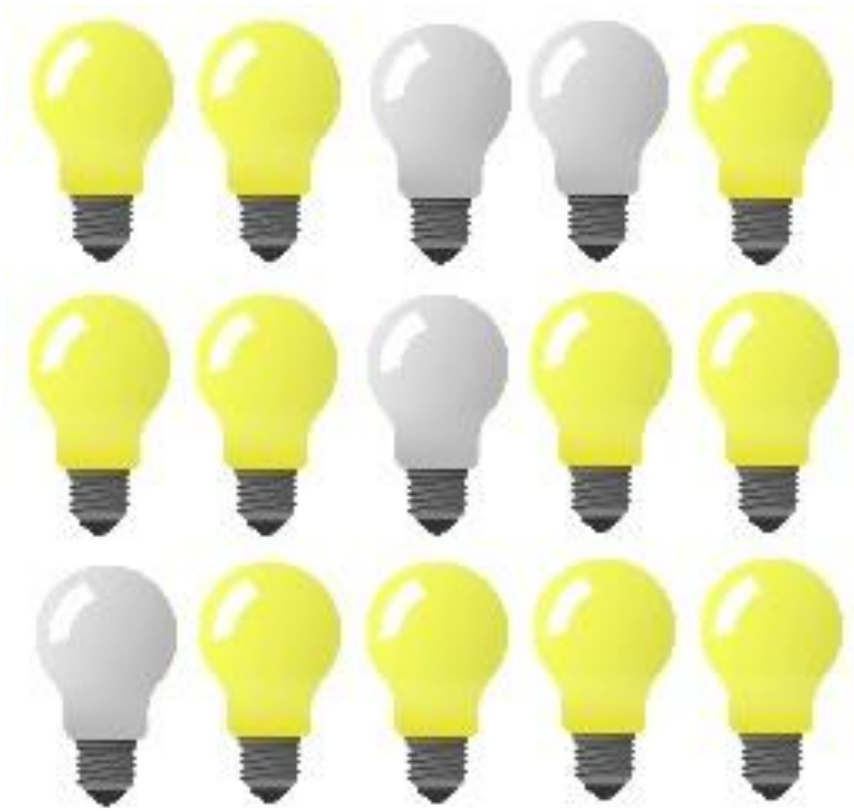
Elena Grigorescu, *Purdue University*

Sidhart Jaggi, *The Chinese University of Hong Kong*

Samson Zhou, *Purdue University*

PURDUE UNIVERSITY

Group Testing



Task: Identify a small set of defects among a larger population using tests

Test = a subset of the items

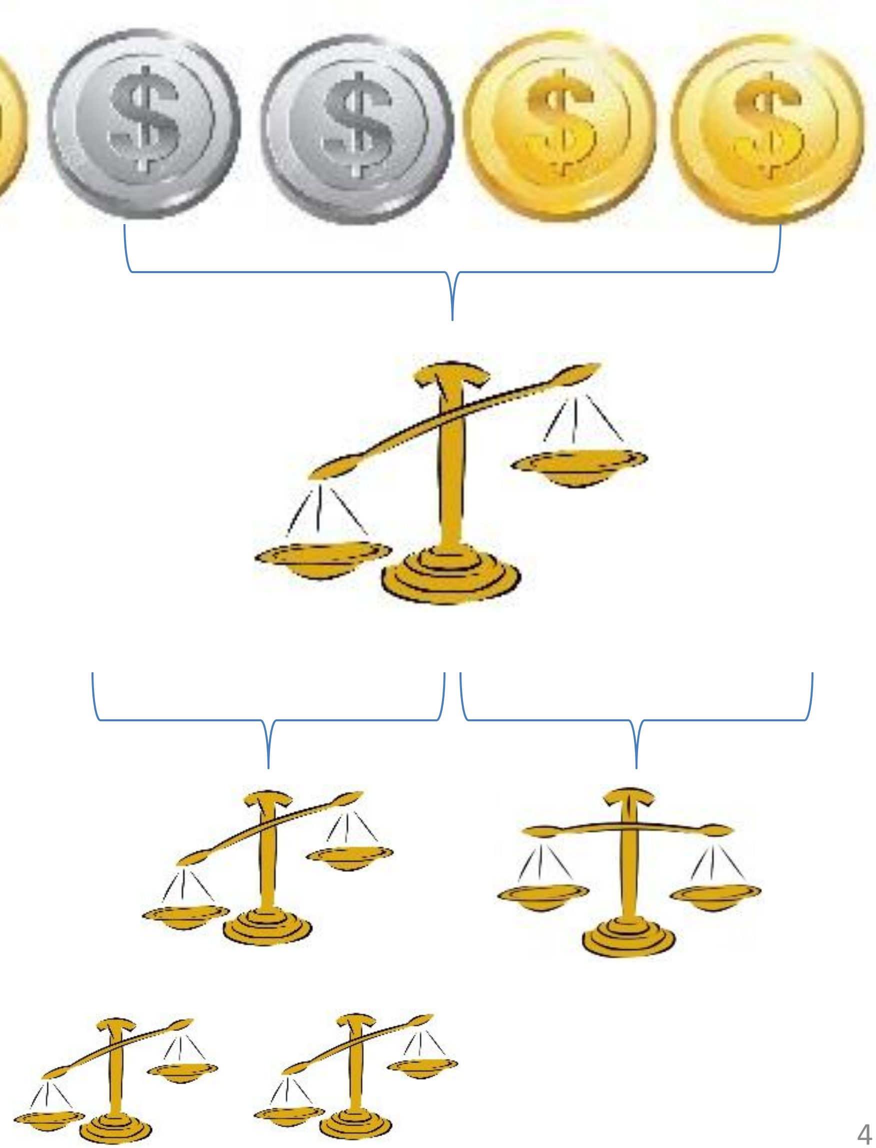
A test is positive if a defective item is included; negative otherwise

Goal: Minimize the number of tests

Background

- Motivation: Identify WW2 draftees with syphillis
- Each test is expensive, but blood may be pooled
- Blood from individual can be used in multiple tests
- How to identify infected individuals?

Adaptive Group Testing



$$T = O(d \log n)$$

$$T = \Omega\left(d \log \frac{n}{d}\right)$$

Non-adaptive Group Testing (NAGT)

Non-adaptive Group Testing (NAGT)

Non-adaptive Group Testing (NAGT)

✓	0	1	0	0	0	0	1	0	0	1	0	0	1
✗	1	1	0	1	0	0	0	1	1	0	0	0	0
✗	0	0	0	0	1	1	1	0	0	1	1	1	0
✗	1	1	1	1	0	0	0	0	0	0	0	1	1
✓	0	0	1	1	0	0	0	1	0	0	1	0	0
✗	0	0	0	0	0	1	1	0	0	0	1	0	1
✗	0	0	1	0	1	0	0	1	0	1	0	0	0
✓	1	0	0	0	0	0	1	1	0	1	1	0	0

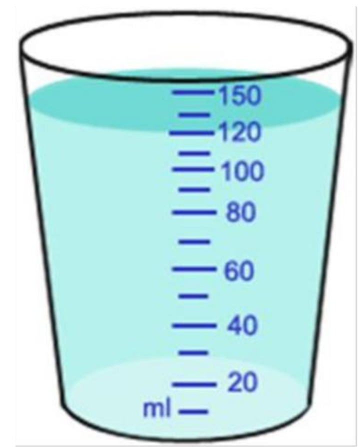
Non-adaptive Group Testing

$$
\begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}
=
\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}
$$

“Classical” NAGT

- Upper Bound: $O\left(d^2 \log n\right)$ (Du, Hwang ‘00)
- Explicit: $O\left(d^2 \log n\right)$ (Porat, Rothschild ’08)
- Lower Bound: $\Omega\left(d^2 \log_d n\right)$ (D’yachkov, Rykov ’82)
- Noisy tests (CheraghchiHKV ’11)
- Efficient Decoding (IndyKRN’10)
- Graph-Constrained (CheraghchiKMS’11)
- Phase-Transitions (Scarlett, Cehver’16)

Real World Limitations



Each item can be included in at most γ tests
(γ -divisible items)

Each test can include at most ρ items
(ρ -sized tests)

Our Results: γ -divisible items

Theorem. Given n items, with d defects: $\Omega\left(\gamma d\left(n/d\right)^{1/\gamma}\right)$ tests are needed in the NAGT, γ -divisible items model.

Theorem. ∃ a randomized algorithm: T = $O\left((\gamma d)\left(\frac{n-d}{\varepsilon}\right)^{1/\gamma}\right)$

Theorem. ∃ a deterministic algorithm: T = $O\left(\frac{d^2\gamma}{\varepsilon}\left(\frac{n\varepsilon}{d^2}\right)^{1/\gamma}\right)$

Our Results: ρ-sized tests

Theorem. Given n items, with d defects:

$$\Omega\left(\frac{n}{\rho}\frac{\log(n/d)}{\log(n/\rho d)}\right)$$ tests are needed in the NAGT, ρ-sized tests model.

Theorem. ∃ a randomized algorithm: T = $O\left(\frac{n}{\rho}\log\left(\frac{n}{\varepsilon}\right)\right)$

Theorem. ∃ a deterministic algorithm: T = $O\left(\frac{n}{\rho}\left(\frac{d^2\log n}{\varepsilon\log(n/\rho)}\right)\right)$

Structure of Talk

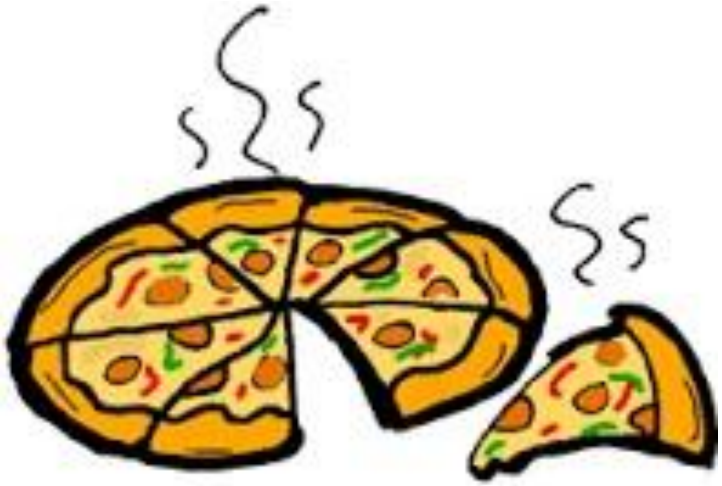
1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Upper Bounds: Randomized

- Consider the γ-divisible items model



Upper Bounds: Randomized

- Recall: Total items tested is at most γn
- Idea: Have each test include roughly $\frac{\gamma n}{T}$ items.
- Columns of test matrix M are uniformly sampled from $\{0,1\}^T$ with weight γ.
- How to decode?

Decoding Algorithm Philosophy

Innocent until
proven guilty

Defective until
proven innocent

Decoding Algorithm

- Algorithm: Marks item *i* non-defective if some test which includes *i* is negative
- Observations:
  - CANNOT incorrectly identify defective items
  - Incorrectly marks non-defective item *i* if all tests which include *i* are positive

Upper Bounds: Randomized

- Recall: Total number of positive tests is at most dγ.
- Probability an item is included only in positive tests: $\binom{d\gamma}{\gamma} / \binom{T}{\gamma}$
- Union bound over all (n-d) non-defective items, so we require:

$$(n-d)\binom{d\gamma}{\gamma} / \binom{T}{\gamma} < \epsilon$$

Upper Bounds: Randomized

- Recall: Total number of positive tests is at most dγ.
- Probability an item is included only in positive tests: $\binom{d\gamma}{\gamma} / \binom{T}{\gamma}$
- Union bound over all (n-d) non-defective items, so we require:

$$(n-d)\binom{d\gamma}{\gamma} / \binom{T}{\gamma} < \epsilon \Rightarrow T > (e\gamma d)\left(\frac{n-d}{\epsilon}\right)^{1/\gamma}$$

Structure of Talk

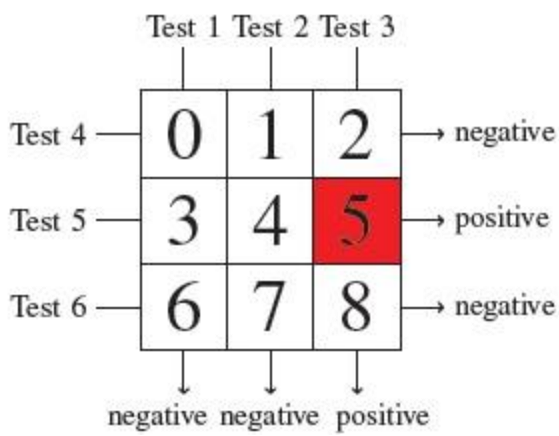
1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Upper Bounds: Deterministic

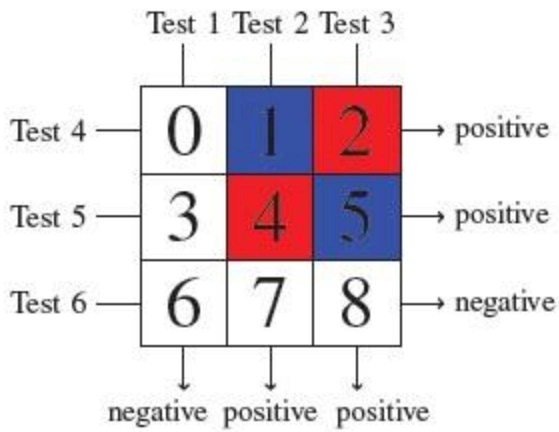
- Intuition: Should be able to “encode” each item so that test outcomes uniquely identify defects
- Idea: Use γ-dimensional hypergrid, represent each item by base-b representation, where $b = n^{1/\gamma}$



If n = 9, γ = 2, d = 1, the above test uniquely determines that item 5 is defective.

Upper Bounds: Deterministic

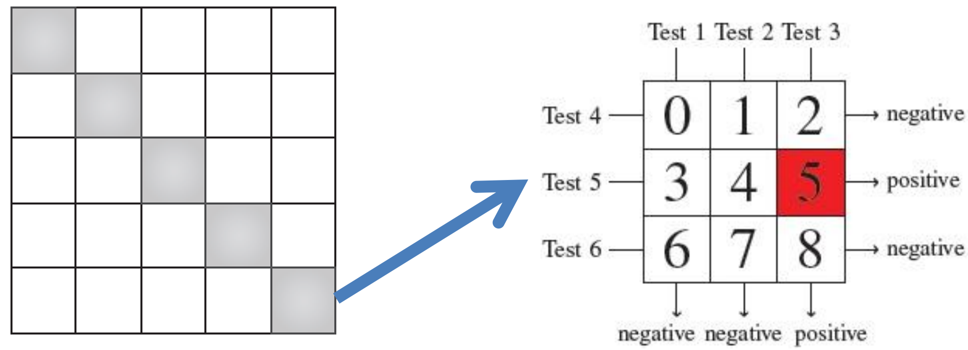
- What if we have multiple defects?



- These tests cannot distinguish whether the red items or the blue items are defective.

Upper Bounds: Deterministic

- What if we have multiple defects?
- Idea: Divide and Conquer!



- Each gray box catches one defective item with high probability

Upper Bounds: Deterministic

- How many blocks are necessary? $\dfrac{d^2}{\varepsilon}$
- Probability that no two defects fall into the same block:

$$
\begin{aligned}
&1\left(1-\frac{1}{cd^2}\right)\left(1-\frac{2}{cd^2}\right)\cdots\left(1-\frac{d-1}{cd^2}\right)\\
&\geq \left(1-\frac{d}{cd^2}\right)^d = \left(1-\frac{1}{cd}\right)^d\\
&\geq 1-\frac{1}{c} = 1-\epsilon \qquad \text{(by Bernoulli's Inequality)}
\end{aligned}
$$

- In summary: $\dfrac{d^2}{\varepsilon}$ blocks, each requiring $\gamma\left(\dfrac{n\varepsilon}{d^2}\right)^{1/\gamma}$ tests, for a total of

$$T = \frac{d^2\gamma}{\varepsilon}\left(\frac{n\varepsilon}{d^2}\right)^{1/\gamma}$$

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Classical NAGT Design Philosophy

0 1 0 0 0 0 1 0 0 1 0 0 1

- Each test gives~1 bit of information

Lower Bounds: γ-divisible items

- Total number of items tested is at most γ n
- "Light" tests: each includes less than $\dfrac{n}{\varepsilon d \log\left(\dfrac{T}{\gamma d}\right)}$ items.
- "Heavy" tests: each includes at least $\dfrac{n}{\varepsilon d \log\left(\dfrac{T}{\gamma d}\right)}$ items.

Lower Bounds: γ-divisible items

- "Light" tests: each includes less than $\dfrac{n}{\epsilon d \log\left(\frac{T}{\gamma d}\right)}$ items.
- "Heavy" tests: each includes at least $\dfrac{n}{\epsilon d \log\left(\frac{T}{\gamma d}\right)}$ items.

$$X \to Y \to \hat{X}$$

$$H(Y) = \sum_{i \in S_1} H(Y_i) + \sum_{i \in S_2} H(Y_i)$$
$$\leq T\left(\frac{\gamma d}{T} + 3\delta\right)\log\left(\frac{T}{\gamma d}\right) + \epsilon\gamma d \log\left(\frac{T}{\gamma d}\right)$$
$$\leq (1+2\epsilon)\gamma d \log\left(\frac{T}{\gamma d}\right)$$

(for appropriate choice of δ).

Lower Bounds: γ-divisible items

$$
\begin{aligned}
H(Y) &= \sum_{i \in S_1} H(Y_i) + \sum_{i \in S_2} H(Y_i) \\
&\le T\left(\frac{\gamma d}{T} + 3\delta\right)\log\left(\frac{T}{\gamma d}\right) + \epsilon\gamma d \log\left(\frac{T}{\gamma d}\right) \\
&\le (1+2\epsilon)\gamma d \log\left(\frac{T}{\gamma d}\right)
\end{aligned}
$$

$$X \to Y \to \hat{X}$$

$$
\begin{aligned}
H(X) &= H(X|\hat{X}) + I(X;\hat{X}) \\
&\le H(\epsilon) + \epsilon \log(|\mathcal{X}| - 1) + H(Y)
\end{aligned}
$$

$$
\log\left(\binom{n}{d}\right) \le -2\epsilon\log\epsilon + \epsilon\log\left(\binom{n}{d}\right) + (1+2\epsilon)\gamma d \log\left(\frac{T}{\gamma d}\right).
$$

Data processing inequality

Fano's inequality

Entropy over all possible combinations of d defects

Lower Bounds: γ-divisible items

$$H(Y) = \sum_{i \in S_1} H(Y_i) + \sum_{i \in S_2} H(Y_i)$$

Implies $T = \Omega\left(\gamma d \left(n/d\right)^{1/\gamma}\right)$

$$\log\left(\binom{n}{d}\right) \leq -2\epsilon \log \epsilon + \epsilon \log\left(\binom{n}{d}\right) + (1+2\epsilon)\gamma d \log\left(\frac{T}{\gamma d}\right).$$

Data processing inequality

Fano's inequality

Entropy over all possible combinations of d defects

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

Structure of Talk

1. Background
2. Achievability: Randomized Construction
3. Achievability: Deterministic Construction
4. Lower Bounds

?? Results for zero-error tests and noisy tests

Further Results

- Noisy tests: Tests can give incorrect result with probability σ.

Theorem: CANNOT recover defectives with probability at least 1- ε for arbitrary ε < 1/2 and γ = o(log n).


рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
kia ora
barka
welalin
tack
spas
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
vinaka
спасибі
blagodaram
kiitos
dankie
nanni
nandri
dhanyavad
bedankt
bayarlalaa
gràçie
hvala
maururu
köszönöm
thank you
gracias
хвала
asante
manana
obrigada
tenki
murakoze
enkosi
akun
dankon
aciū
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dékuji
obrigado
sagolun
sukriya
kop khun krap
taiku
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
rahmet
kam sah hamnida
rahmat
terima kasih
diolch
dhanyavadagalu
shukriya
merce
мерси
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci

Questions?