Noisy Group Testing and Boolean Compressed Sensing

Venkatesh Saligrama

Boston University

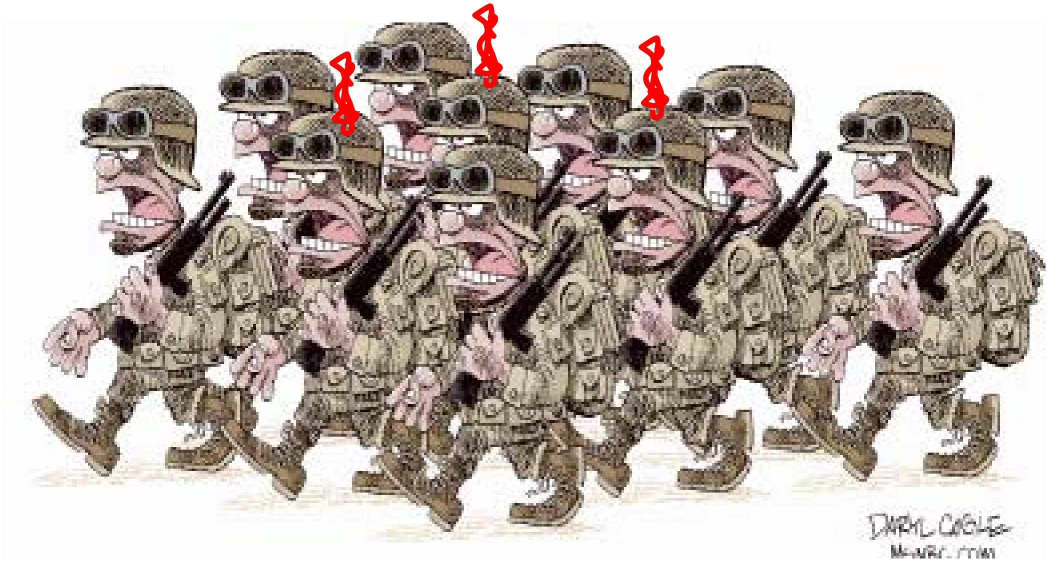
Collaborator: George Atia

Outline

- Examples …
- Problem Setup
  - Noiseless Problem
    - Average error,
    - Worst case error
    - Approximate reconstruction
  - Noisy Problem
    - Additive Noise (False Alarms)
    - Dilution Effect (Misses)
- Background Material
- Information-Theoretic Analysis: Achievability & Converse
  - Tradeoffs between #tests, #defectives, Noise

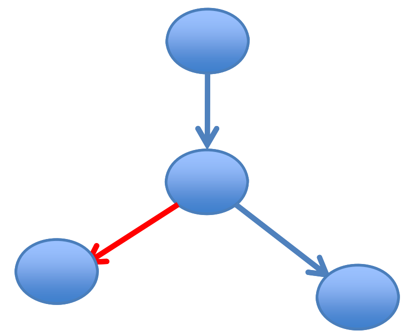
What is Group testing?

- Few soldiers in a large population have disease
  - Detect by testing pooled blood samples
    - Compressed Sensing in 1940s!!



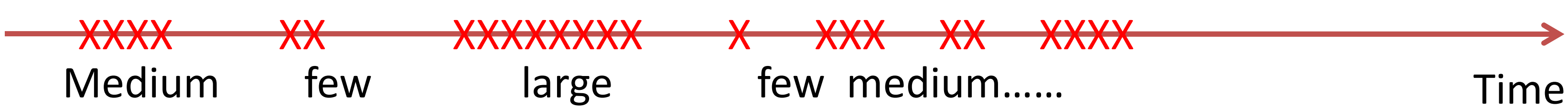
Applications

- DNA Screening (clones) [Du,Hwang'00]
- Streaming [Gilbert-Strauss 08]
  - Telephone Calls
- Compressed Sensing [Muthukrishnan 05]



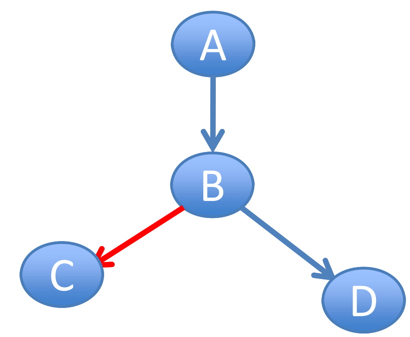
- Congested IP Link [Nguyen-Thiran07]
- Spectrum Violation in Cognitive Radio [Atia-S-S 08]

MAC layer model of packet drops



The group testing problem

- Given
  - N items: $R_1, R_2, R_3, \ldots, R_N \in \{0,1\}$
  - K defectives: $R_m = 1$
    - $j \in S \subset \{1, 2, \ldots N\}$



- Test Matrix: $C = [X_{mn}]$,
  - $X_{mn} = 1 \rightarrow$ Put nth item in pool(test) m
- Output (Y_m) – mth test Outcome: +ve or –ve
  - Channel(noiseless)

$$Y_m = \bigvee_{n=1}^{N} X_{mn} R_n \longrightarrow Y = C\,R$$

Illustrative Example

Items \ Tests	1	2	3 …												T
Y	0	0	1	0	1	1	0	0	0	0	1	0	1	0	0
1	1	0	0	0	0	1	1	0	1	0	1	0	0	0	0
2 (Defectives)	0	0	1	0	1	1	0	0	0	0	0	0	1	0	0
(Defectives)	0	0	1	0	1	1	0	0	0	0	1	0	1	0	0
⋮															
N	1	1	0	1	0	0	0	1	0	1	0	1	0	0	0

- Non-Adaptive vs Adaptive
- Algorithm: Greedy (Matching algorithm)

Noiseless Problem

- Given, Sample size N, and Defective size K,
  - Find a Test Matrix $C \in \{0, 1\}^{N \times T}$
    - Small Misclassification Error
    - #Tests = T → minimum

- Misclassification
  - Defectives indexed by $s \in \mathcal{D} = \{1, 2, \ldots \binom{N}{K}\}$
  - Decoder $g : Y^T \rightarrow \mathcal{D}$

Indicator function

  - Pointwise Error: $\lambda_s = \mathcal{I}(g(Y^T) \neq s | s)$
  - other errors

Problem Statements (Noiseless)

- Our Focus: Does a design matrix $C \in \{0, 1\}^{N \times T}$ exist?
  - Average Error is small $\lambda_{avg}(C) = \frac{1}{\binom{N}{K}} \sum_s \lambda_s(C) \leq \epsilon$
  - Worst-case error is small $\lambda_{\max}(C) = \max_s \lambda_s(C) \leq \epsilon$
  - Asymptotic → Approaches Zero with large K and N.
  - Distortion: $\lambda_s(C) = \mathcal{I}(\underbrace{d(g(Y^T), s)}_{\text{Distance function}} > \alpha K \mid s)$

Noisy Case 1

- Additive noise

$$Y_m = \bigvee_{n=1}^{N} X_{mn} R_n \bigvee W_m$$

W_m Bernoulli(q), m=1, 2, …T

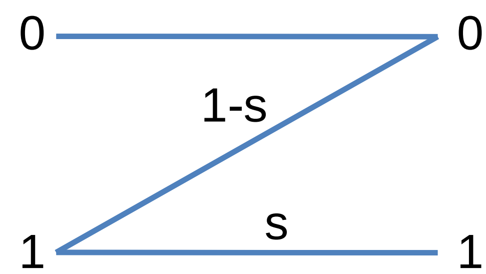
Motivation:

- False alarms of tests
- Background losses (Wireless)

Noisy case 2: Dilution effect

Situation when item is diluted in the pool

$$Y_m = \bigvee_{n=1}^{N} \mathcal{Z}(X_{mn} R_n)$$



- Dilution effects in blood tests or DNA screening
- Probabilistic adversarial transmission (Asilomar'08)
- Link Losses [Nguyen-Thiran07]

Allowable transmission pattern 0 1 1 0 1 0 0 0 1 0 ……

Dilution

1-s s s 1-s

Actual transmission pattern 0 0 1 0 1 0 0 0 0 0 ……

Problem

- Does there exist a matrix C

$$\bar{\lambda}_s(C) = E(\lambda_s(C)) = \text{Prob}(g(Y^T) \neq s|s))$$

Averaged over noise

  - Small misclassification error
  - worst-case, average-case, distortion, asymptotics

Prior Work: Noiseless Cases

- Adaptive and non adaptive group testing (Du, Hwang'2000)

- Non Adaptive
  - Superimposed codes (Kautz and Singleton'64)
  - Deterministic designs (Dyachkov and Rykov'83) (Ruszinko'94)(Erdos'85)(Ngyuen'88)(Porat'08)
  - Random Designs (Dyachkov'76,'82)(Sebo'85)(Macula'96)
  - Compressed sensing and approximate identification(Gilbert'08)
  - Two-Stage Disjuntive Testing: (Berger-Levenshtein 2002)

Our Approach/Contribution

- Random Coding perspective
  - Information theoretic relationship
    - Misclassification vs. Existence of Random Matrix
  - Mutual Information Formula for different problems
    - Meets existing bounds
  - Extensions to new problems → Noisy Case

Main Result

- Theorem (average error → 0 asymptotically) if:

$$T = \max_i \left\{ \frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{(i)}; X_{(K-i)}, Y)} \right\}$$

sufficiency

$$T \geq \frac{\log \binom{N}{K}}{I(X_{(K)}, Y)}$$

Necessity

- X_{mn} generated i.i.d. Bernoulli p
- $X_{(K)}$ corresponds to collection associated with K defective items
- $X_{(i)}$ subset of i defective items in K (that are mis-classified)
- Necessity: FANO Bound

Avg Error Interpretation

$$T = \max_i \left\{ \frac{\log \binom{N-K}{i}\binom{K}{i}}{I(X_{(i)}; X_{(K-i)}, Y)} \right\}$$

#ways i of the K items are mis-classified

Amount of Info if K-i items revealed

Noiseless Case Computation

$$I(X_{(1)}; X_{(K-1)} Y) = H(Y/X_{(K-1)}) - H(Y/X_{(K)})$$

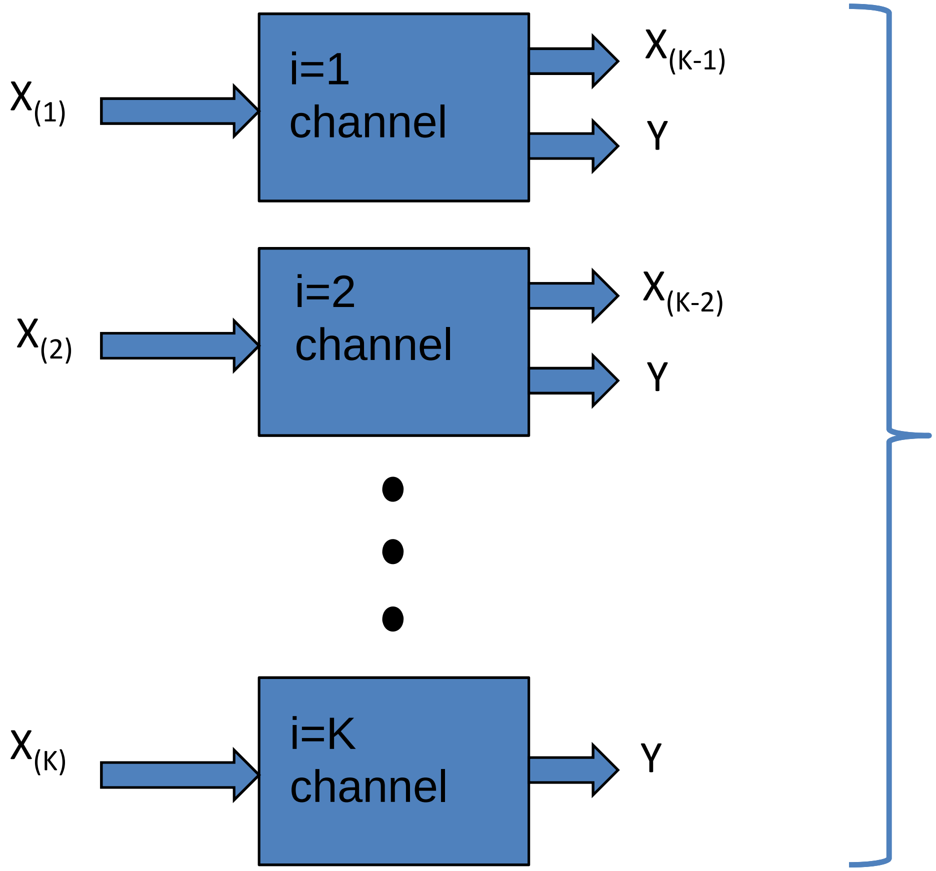
0

Prob($X_{(K-1)}$ = 0)

$(1-p)^{k-1}$ H(p)

$$T = O(K \log N)$$

Noiseless- average P_e



- Typical Set Decoding
- 1 error typical, 2 errors typical ….
- Main Difficulty:
  - Channel Coding: Another Codeword is Typical is Independent of Test Output Y.
  - Here: Another Collection can be overlapping with True Collection

- Errors:
  - •True collection not typical
  - •Another collection typical

$$P_e \leq P(E_0^c) + \sum_i P(E_i)$$

$P(E_i)$: Probability that a set which differs from true coalition in exactly i users is jointly typical

Summary of Results

- N items; K defectives; T Pools/Tests

$T = O(K \log N)$ Noiseless- average P_e

$T = O(K^2 \log N)$ Noiseless- Max P_e (exact reconstruction)
Compare with CS

$T = O(K \log N)$ With distortion (approximate reconstruction)

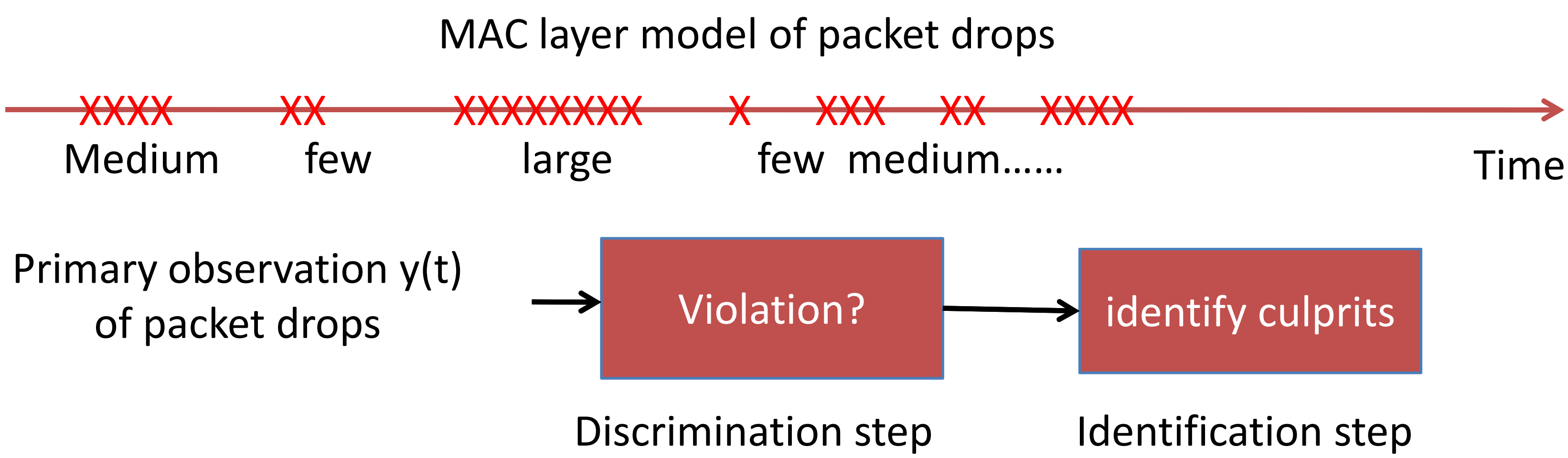
$T = O\left(\frac{K \log N}{\log 1/q}\right)$ Additive Noise (False alarms)

$T = O(K \log N)$ Dilution Effects

Conclusions

- Random Coding Analysis of Group Testing
- Mutual Info Expression
- Easy to Compute
- Extensions to Noisy Cases

Identity through interference fingerprints



Our Data is: Collision/No collision in the different time slots

1 0 0 1 1 0 0 1 1 1 0 0 0 1 0…

Atia, Sahai and Saligrama, Dyspan'08
Atia, Saligrama and Sahai, Asilomar '08

Theorem

For Constant Time till conviction T_c and sparse number of culprits (K) we can support as many users N with throughput of order p*N (i.e. fixed utilization)