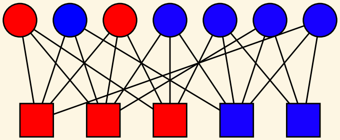
Group Testing

Amin Coja-Oghlan

Goethe University Frankfurt

based on joint work with Oliver Gebhard, Max Hahn-Klimroth, Phillip Loick

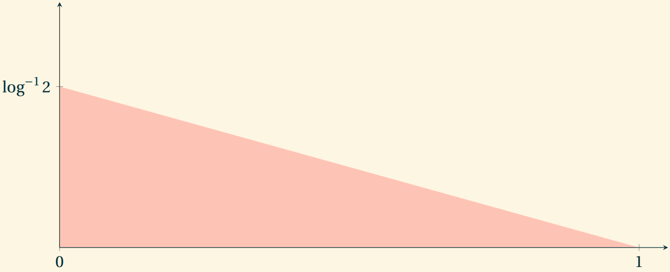
Group testing



The problem [D43,DH93]

- n =population size, $k = n^{\theta}$ = #infected, m = #tests
- all tests conducted in parallel [non-adaptive]
- how many tests are necessary…
- …information-theoretically?
- …algorithmically?

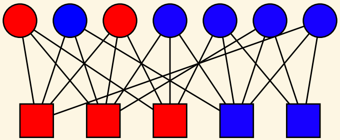
Information-theoretic lower bounds



- if $k \sim n^\theta$ we need

$$2^m \geq \binom{n}{k} \qquad \Rightarrow \qquad m \geq \frac{1-\theta}{\log 2} \cdot k \log n$$

Random hypergraphs



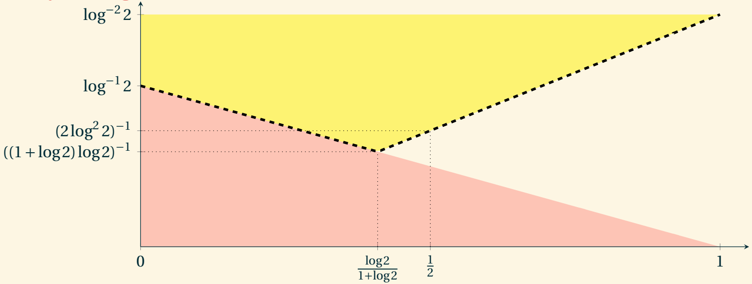
A randomised test design [JAS16,A17]

- a random Δ-regular Γ-uniform hypergraph with

$$\Delta \sim \frac{m \log 2}{k}, \qquad \Gamma \sim \frac{n \log 2}{k}$$

- the choice of Δ, Γ maximises the entropy of the test results

Random hypergraphs



Theorem

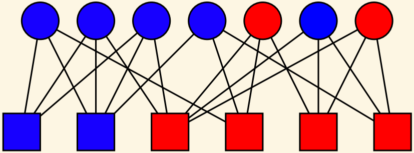
Let

$$m_{\mathrm{rnd}} = \max\left\{\frac{1-\theta}{\log 2}, \frac{\theta}{\log^2 2}\right\} k \log n \qquad \text{where} \quad k \sim n^{\theta}$$

The inference problem on the random hypergraph

- is insoluble if $m < (1-\varepsilon) m_{\mathrm{rnd}}$ [JAS16]
- reduces to hypergraph VC if $m > (1+\varepsilon) m_{\mathrm{rnd}}$ [COGHKL19]

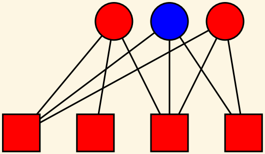
Greedy algorithms



`DD`: Definitive Defectives [ABJ14]

- declare all individuals in negative tests uninfected
- check for positive tests with just one undiagnosed individual
- declare those individuals infected
- declare all others uninfected
- $\rightsquigarrow$ *may produce false negatives*

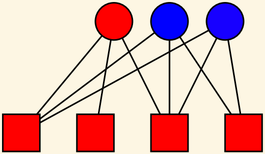
Greedy algorithms



DD: Definitive Defectives [ABJ14]

- declare all individuals in negative tests uninfected
- check for positive tests with just one undiagnosed individual
- declare those individuals infected
- declare all others uninfected
- $\rightsquigarrow$ *may produce false negatives*

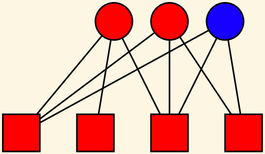
Greedy algorithms



`DD`: Definitive Defectives [ABJ14]

- declare all individuals in negative tests uninfected
- check for positive tests with just one undiagnosed individual
- declare those individuals infected
- declare all others uninfected
- $\rightsquigarrow$ *may produce false negatives*

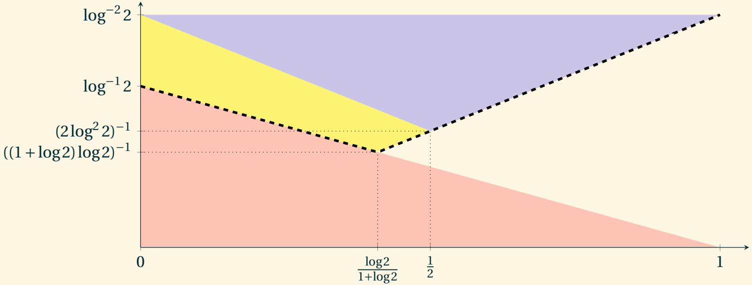
Greedy algorithms



`SCOMP`: greedy vertex cover [ABJ14]

- declare all individuals in negative tests uninfected
- check for positive tests with just one undiagnosed individual
- declare those individuals infected
- greedily cover the remaining positive tests
- $\rightsquigarrow$ *may produce false positives/negatives*
- *Conjecture:* `SCOMP` strictly outperforms `DD` [ABJ14]

Greedy algorithms



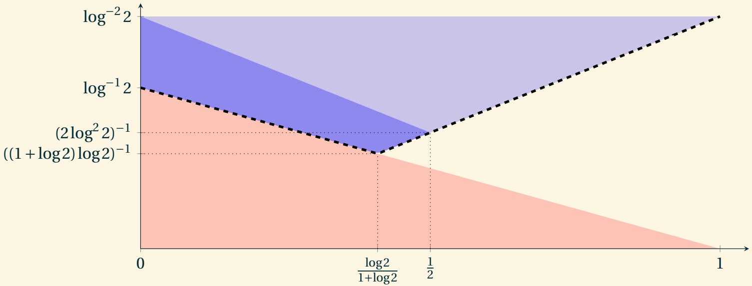
Theorem [ABJ14,COGHKL19]

Let

$$m_{\mathrm{DD}} = \frac{\max\{1-\theta,\theta\}}{\log^2 2} k \log n$$

- if $m > (1+\varepsilon) m_{\mathrm{DD}}$ then both `DD` and `SCOMP` succeed
- if $m < (1-\varepsilon) m_{\mathrm{DD}}$ then both of them fail

The SPIV algorithm

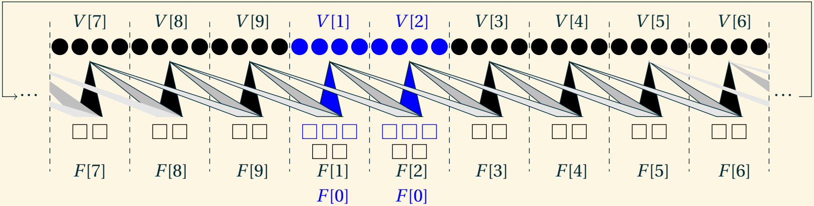


Theorem [COGHKL20]

There exist a test design and an efficient algorithm SPIV that succeed w.h.p. for

$$m \sim m_{\mathrm{rnd}} = \max\left\{\frac{1-\theta}{\log 2}, \frac{\theta}{\log^2 2}\right\} k \log n$$

The SPIV algorithm

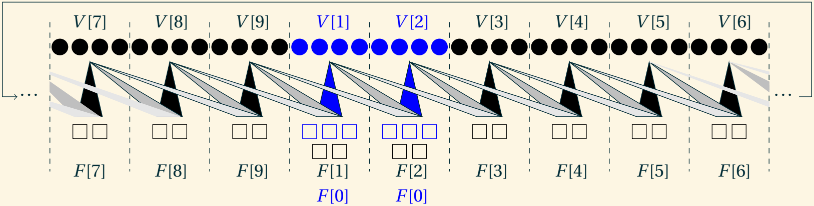


Spatial coupling

- a ring comprising $1 \ll \ell \ll \log n$ compartments
- individuals join tests within a sliding window of size $1 \ll s \ll \ell$
- extra tests at the start facilitate DD

inspired by low-density parity check codes *[KMRU10]*

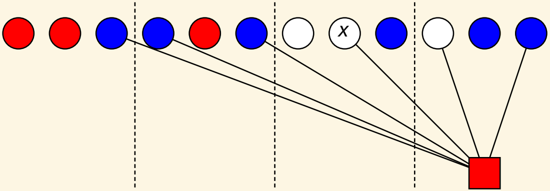
The SPIV algorithm



The algorithm

1. run DD on the s seed compartments
2. declare all individuals that appear in negative tests uninfected
3. tentatively declare infected k/ℓ individuals with max score W_x
4. combinatorial clean-up step

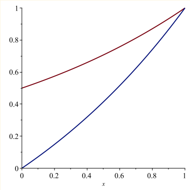
The SPIV algorithm

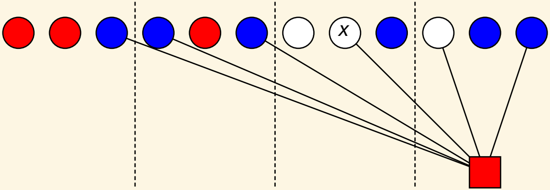


Unexplained tests

- let $W_{x,j}$ be the number of 'unexplained' positive tests $j-1$ compartments to the right of x

The SPIV algorithm

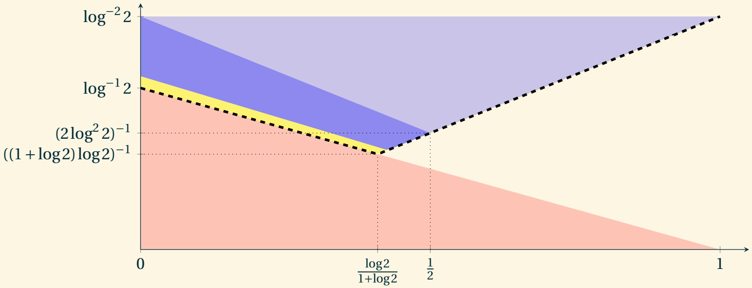




Unexplained tests

- if x is infected, then $W_{x,j} \sim \mathrm{Bin}(\Delta/s, 2^{j/s-1})$
- if x is uninfected, then $W_{x,j} \sim \mathrm{Bin}(\Delta/s, 2^{j/s} - 1)$

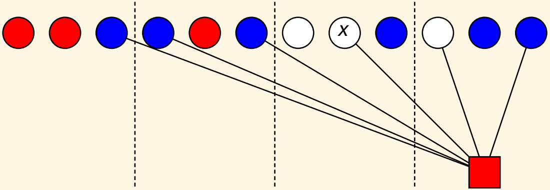
The SPIV algorithm



The score: first attempt

- just count unexplained tests
- we find the large deviations rate function of $\sum_{j=1}^{s-1} W_{x,j}$
- unfortunately, we will likely misclassify $\gg k$ individuals

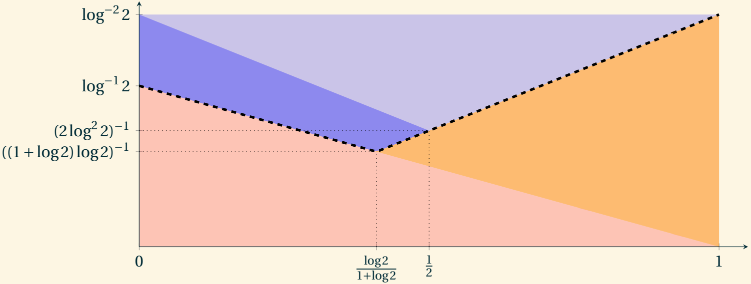
The SPIV algorithm



The score: second attempt

- consider a weighted sum $W_x = \sum_{j=1}^{s-1} w_j W_{x,j}$
- Belief Propagation $\rightsquigarrow$ optimal weights $w_j = -\log(1 - 2^{-j/s})$
- only $o(k)$ misclassifications

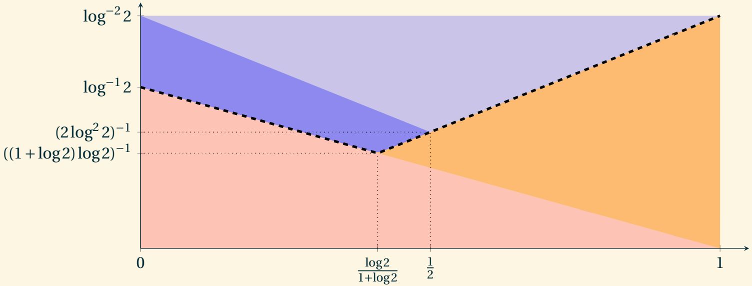
A matching lower bound



Theorem [COGHKL19]

Non-adaptive group testing is information-theoretically impossible with $(1-\varepsilon)m_{\mathrm{rnd}}$ tests.

A matching lower bound



Proof strategy

- *Dilution:* it suffices to consider $\theta = 1 - \delta$
- *Regularisation:* optimal designs are approximately regular
- *Positive correlation:* probability of being disguised [MT11,A18]
- *Probabilistic method:* disguised individuals likely exist

A matching lower bound

Dilution

- assume that for *some* $\log(2)/(1+\log(2)) < \theta < 1$ we get by with

$$m < (1-\varepsilon)\frac{\theta}{\log^2 2}k\log n$$

- then this improvement extends to *all*

$$\frac{\log(2)}{1+\log(2)} < \theta < 1$$

- just add a suitable number of healty dummies
- hence we may assume $\theta = 1-\delta$

A matching lower bound

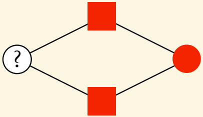
Regularisation

- we may assume that there are no tests of size greater than

$$\frac{n}{k}\log n$$

- $\Rightarrow$ no more than $\frac{n}{\log n}$ individuals have degree more than $\log^3 n$

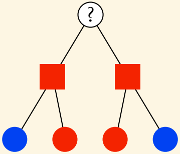
A matching lower bound



Positive correlation

- assume $\theta > 1-\delta$ for a small $\delta > 0$
- FKG inequality $\Rightarrow$ it's a bad idea to create short cycles
- good designs locally resemble a (Δ,Γ)-regular tree

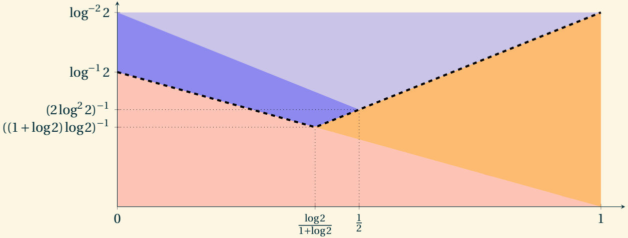
A matching lower bound



Probablistic method

- call an individual x *disguised* if every test $a \in \partial x$ contains another individual $y \neq x$ that is infected
- many disguised healthy *and* infected individuals
- therefore, there are several solutions

Adaptive group testing



Beating the lower bound

- tests are conducted in several stages
- *Goal:* to minimise the number of tests and of stages
- a 3-stage design and algorithm are known with [S19]

$$m \sim \frac{1-\theta}{\log 2} k \log n$$

An optimal 2-stage design

Stage 1

- use the spatially coupled test design with

$$m \sim \frac{1-\theta}{\log 2} k \log n, \qquad \Delta \sim (1-\theta)\log n, \qquad \Gamma \sim \frac{n \log 2}{k}$$

- apply Steps 1–3 of `SPIV`
- drop the clean-up step

An optimal 2-stage design

Stage 2

- test each individual that Stage 1 deems infected separately
- to the rest apply the random hypergraph design and DD with

$$m' = k, \qquad \Delta' = \lceil 10 \log n \rceil$$

- $\rightsquigarrow O(k)$ tests in total

An optimal 2-stage design

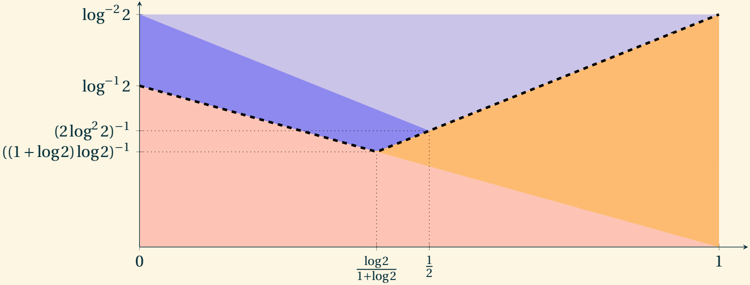
Theorem [COGHKL20]

There exist a 2-stage test design and an efficient inference algorithm with

$$m \sim \frac{1-\theta}{\log 2} k \log n.$$

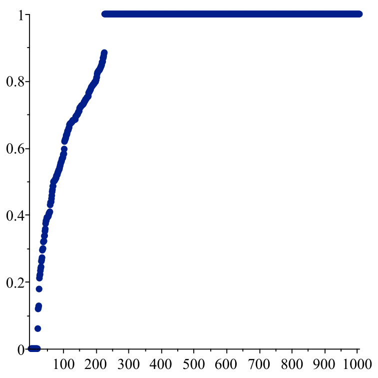
Matches the counting lower bound.

Contributions



- optimal efficient non-adaptive algorithm `SPIV`
- matching information-theoretic lower bound
- optimal two-round adaptive algorithm

Practical group testing



- in wet lab one should assume $k = \Theta(n)$
- non-adaptive testing impossible [A19]
- Belief Propagation leads to promising multi-stage schemes

Open problems

- optimal adaptive designs in the linear case
- combinatorial group testing
- further applications of spatial coupling
- *practical group testing*

`https://arxiv.org/abs/1911.02287`