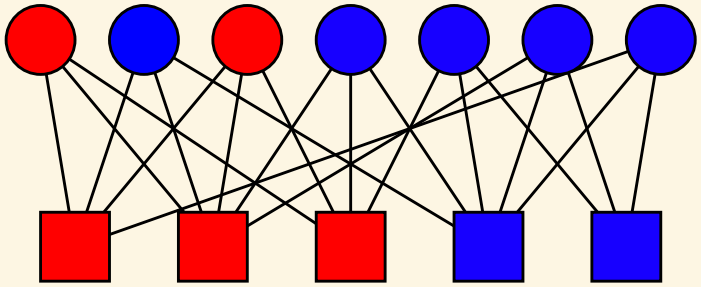
Optimal Group Testing

Amin Coja-Oghlan

Goethe University Frankfurt

joint work with Oliver Gebhard, Max Hahn-Klimroth, Philipp Loick

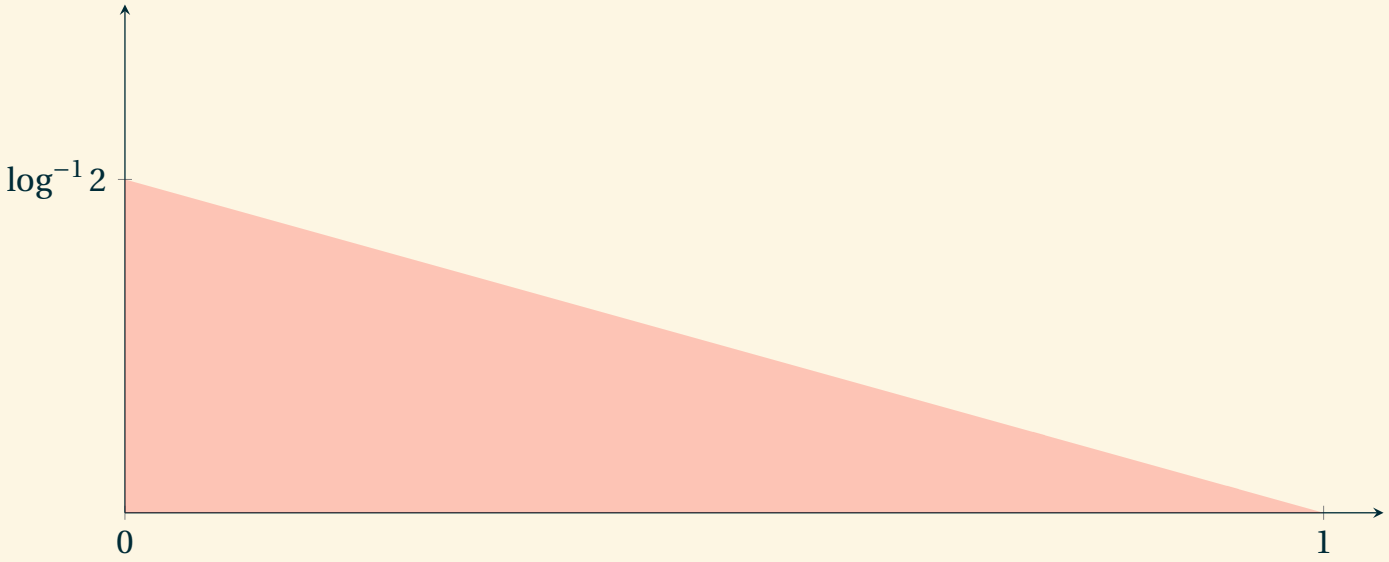
The problem



Group testing [D43,DH93]

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

Information-theoretic lower bounds

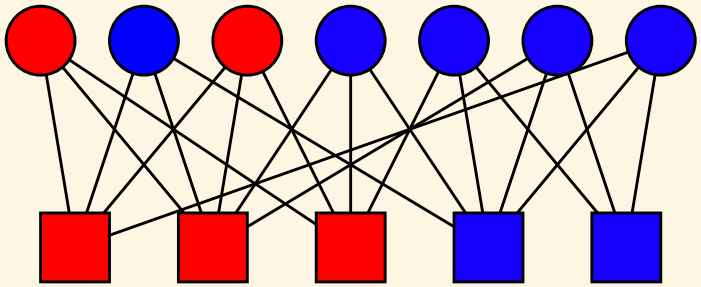


- if $k \sim n^\theta$ we need

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

- if $k = \Theta(n)$ we inevitably need $m = n$ tests [A18]

Random hypergraphs



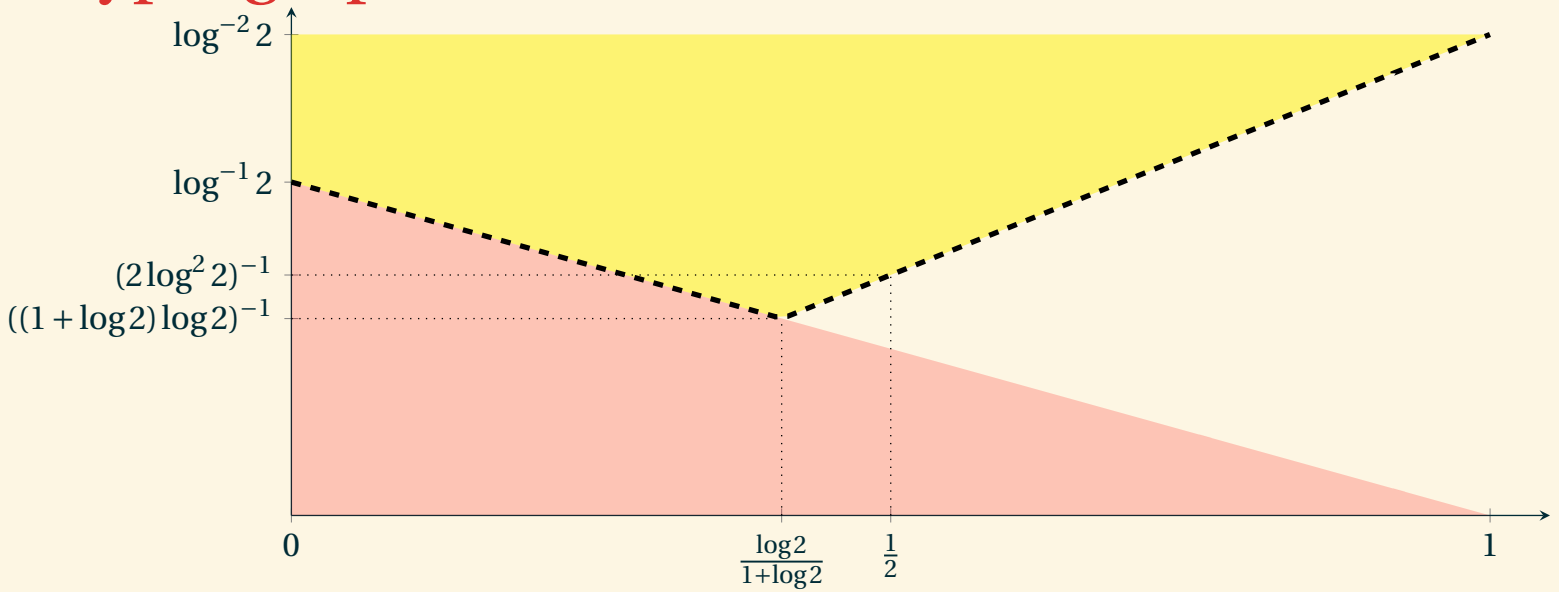
A randomised test design [JAS16,A17]

- 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

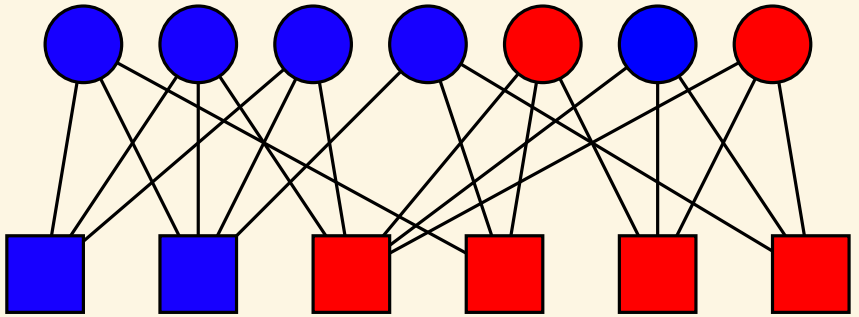
Let

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

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]

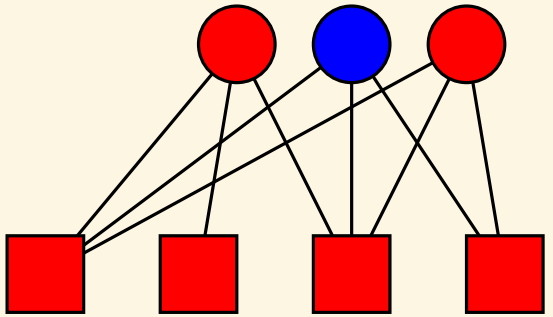
Greedy algorithms



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*

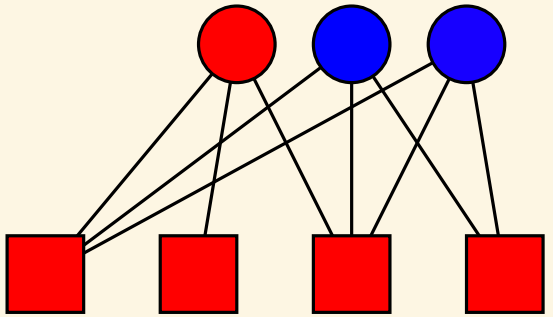
Greedy algorithms



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*

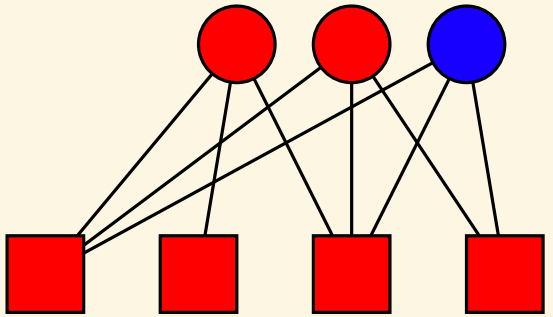
Greedy algorithms



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*

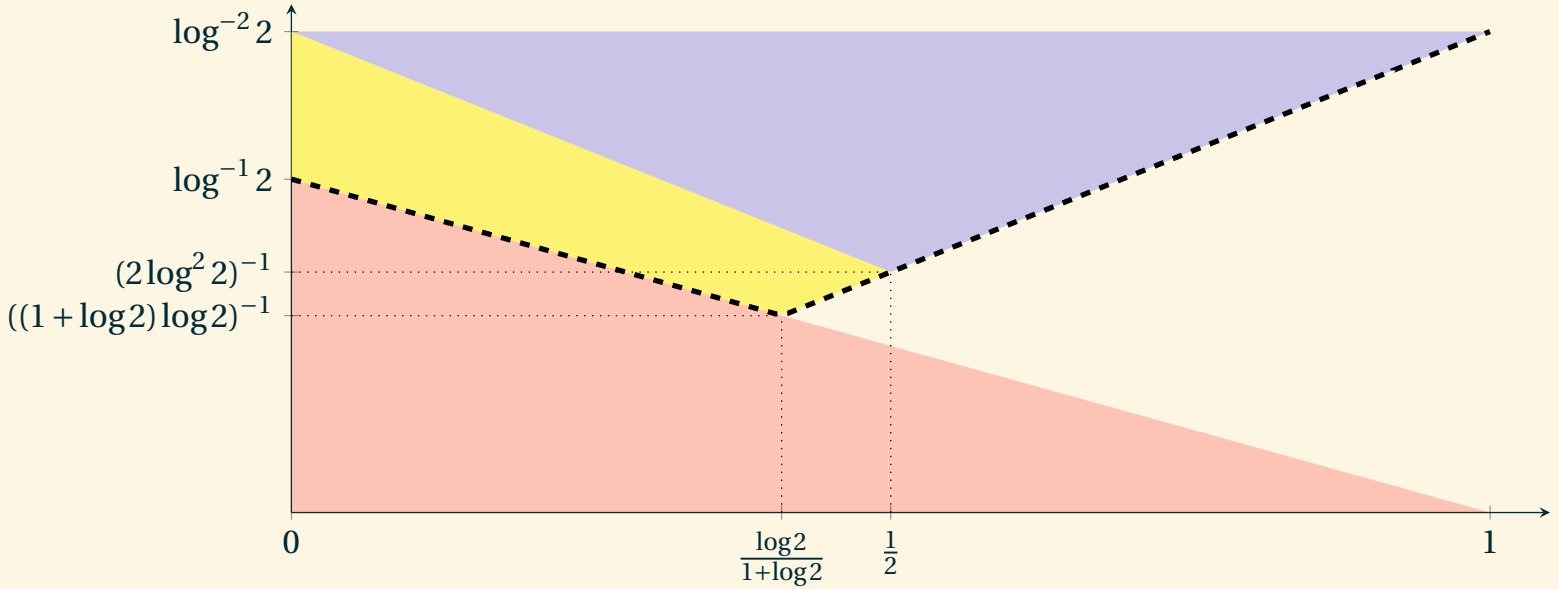
Greedy algorithms



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



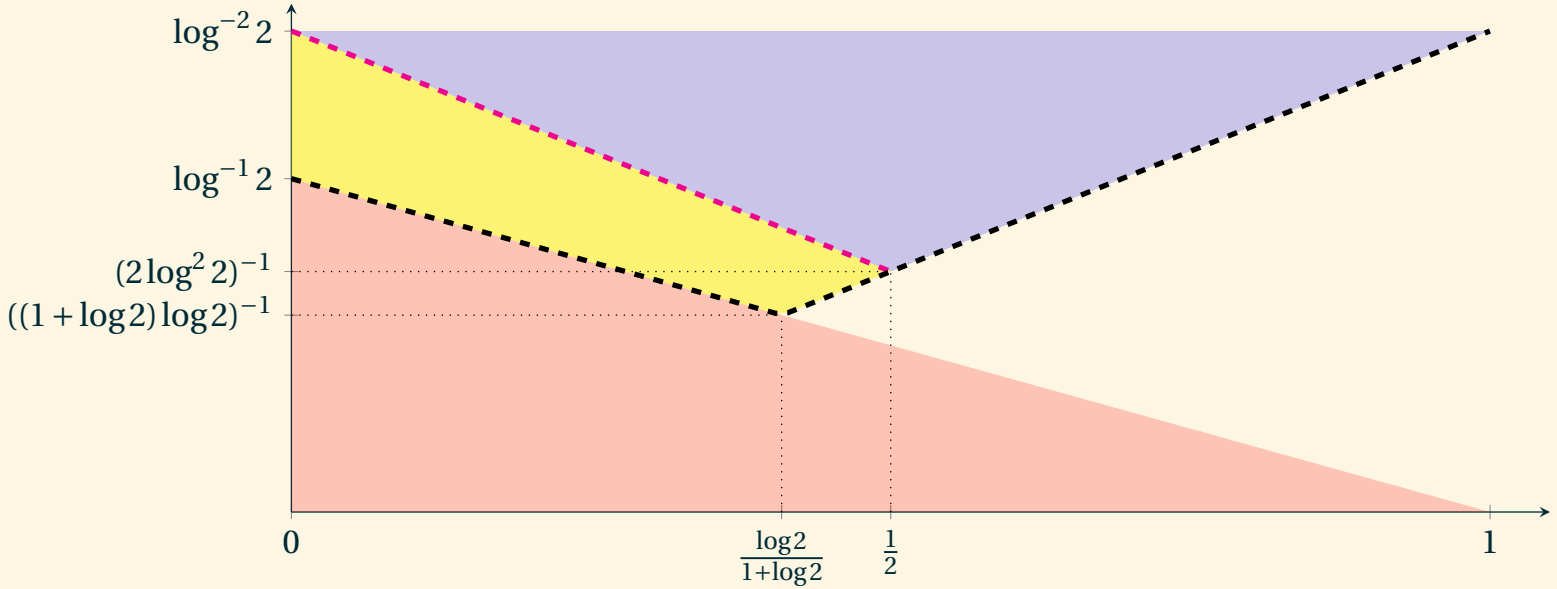
Theorem

Let

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

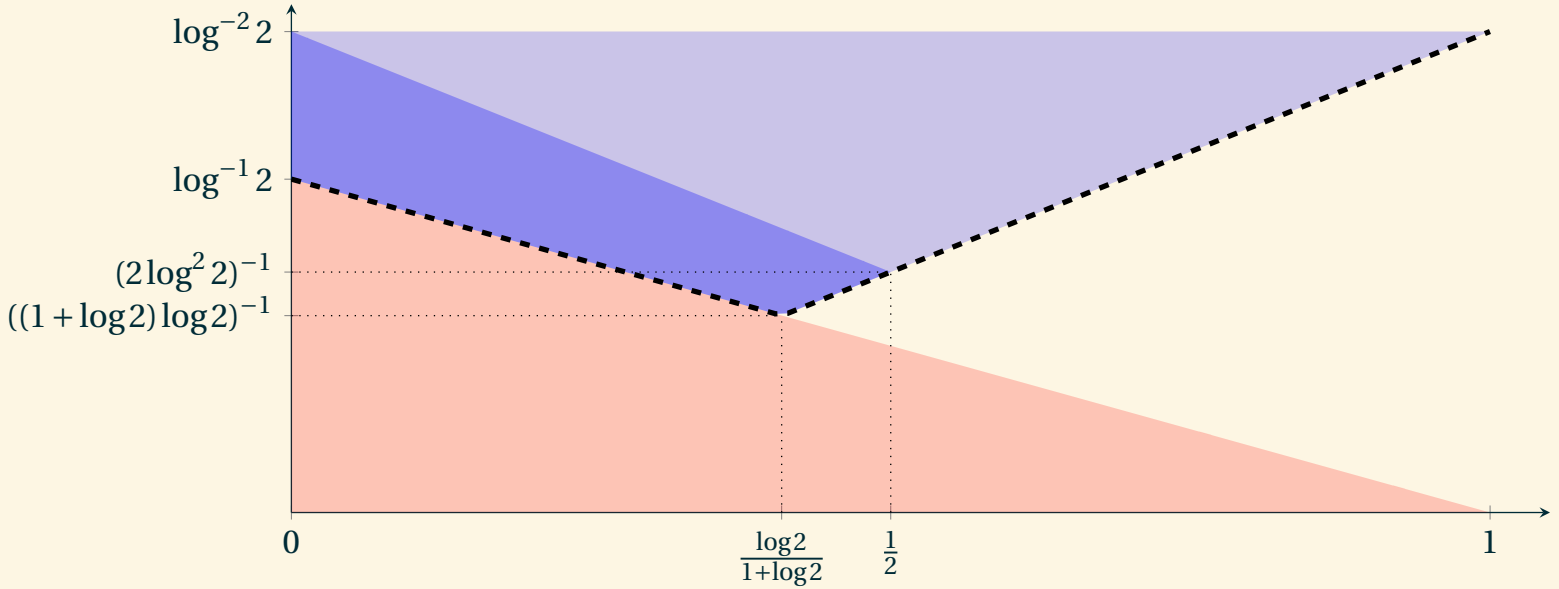
- if $m > (1+\varepsilon) m_{\mathrm{DD}}$, then both DD and SCOMP succeed [ABJ14]
- if $m < (1-\varepsilon) m_{\mathrm{DD}}$, then both of them fail [COGHKL19]

Prior work: summary



- the counting bound
- the cavity method, two-stages, FKG lower bound [MTT07]
- greedy algorithms: positive [ABJ14]
- greedy algorithms: negative [COGHKL19]

The SPIV algorithm



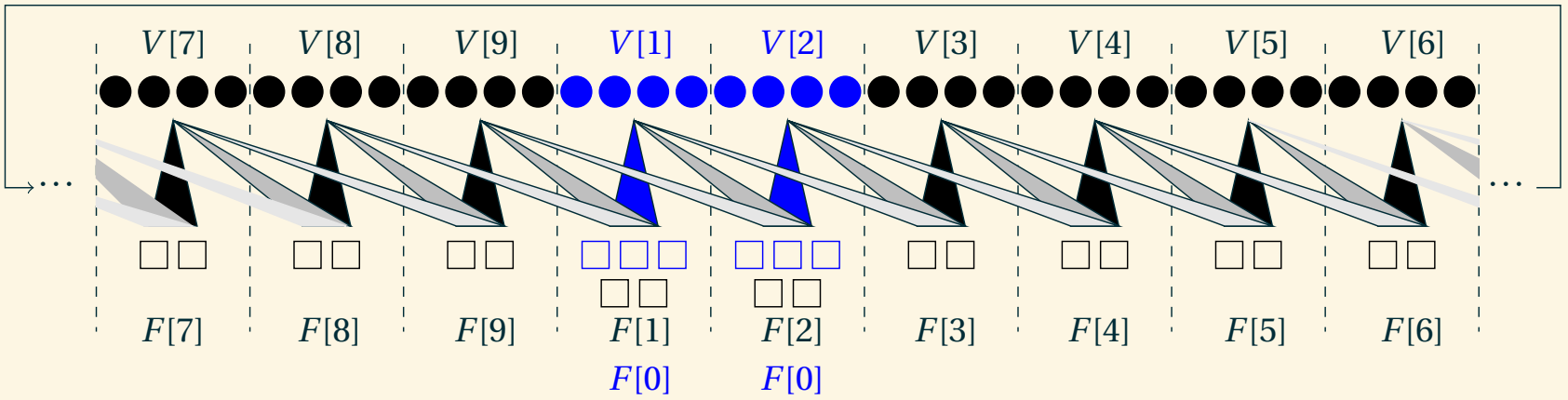
Theorem

[COGHKL19]

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

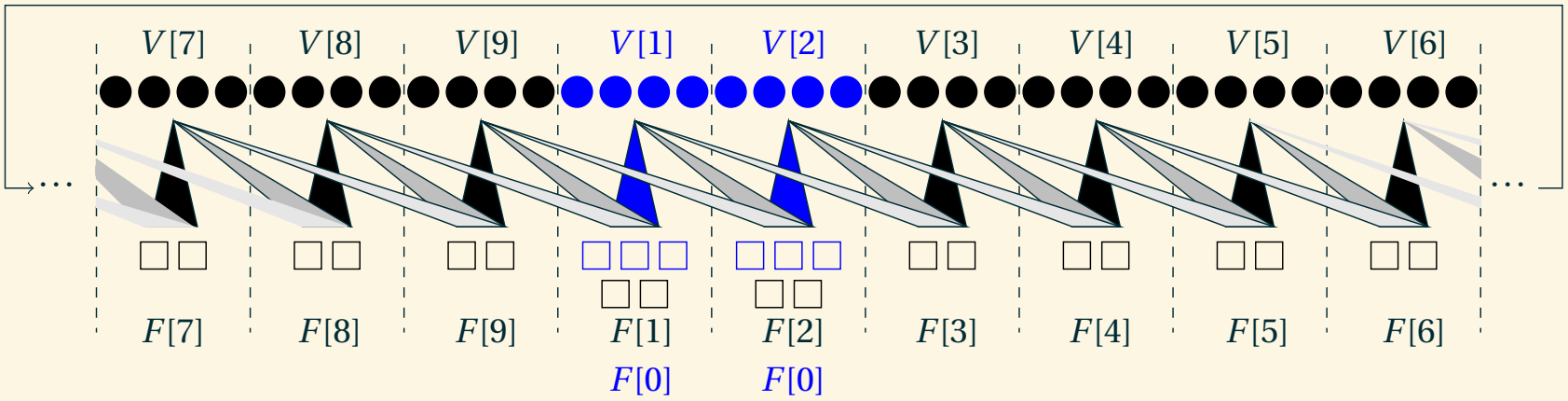


Spatial coupling

- 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]*

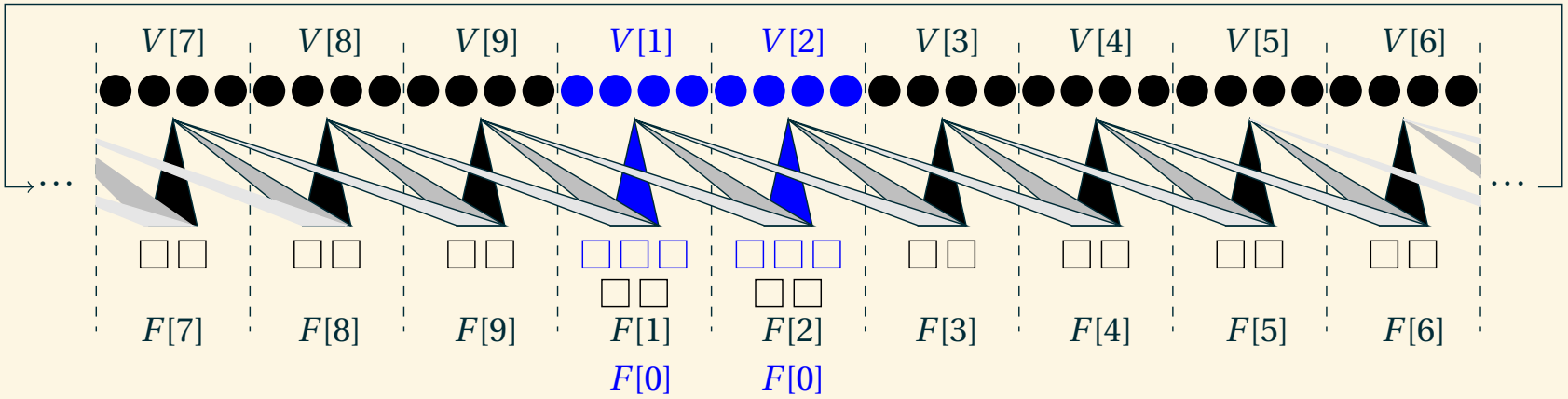
The SPIV algorithm



Spatial coupling

- low-density parity check codes [KMRU10]
- compressed sensing [KMSSZ11,DJM13]
- quantitative group testing [ZKMZ13]
- spatial coupling as a proof technique [GMU12]

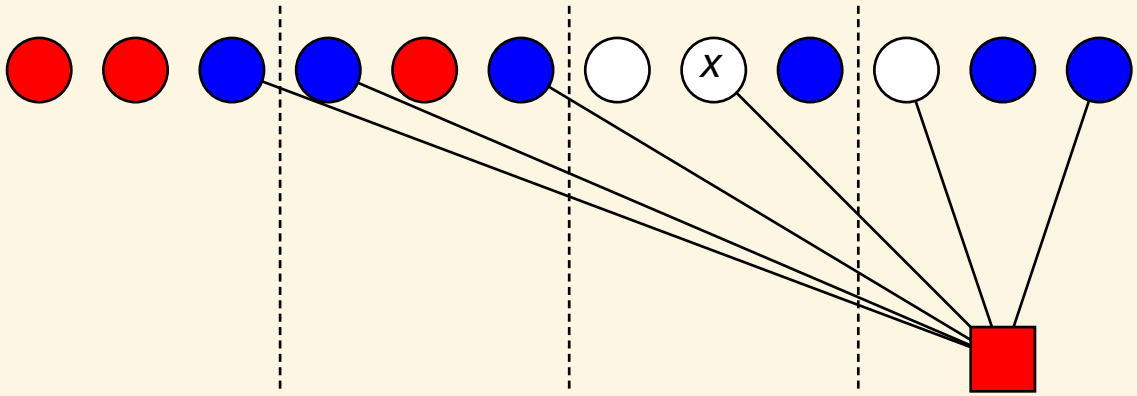
The SPIV algorithm



The algorithm

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

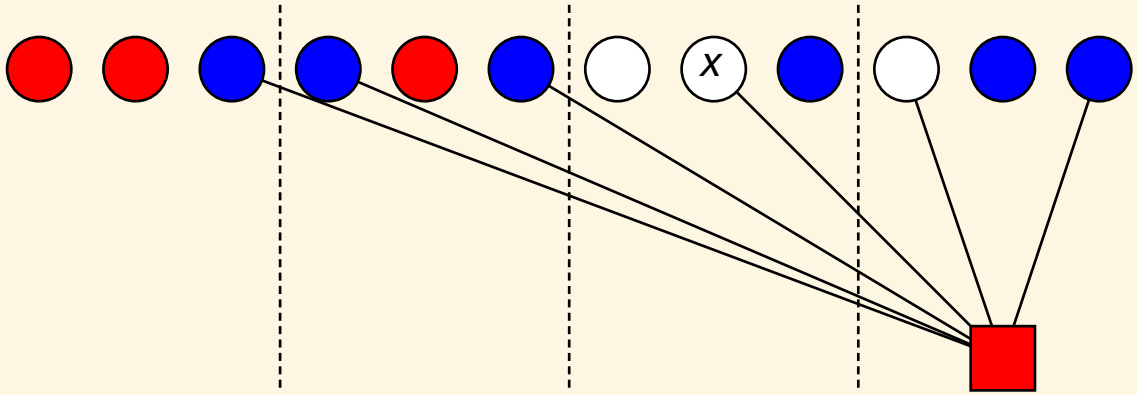
The SPIV algorithm



Unexplained tests

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

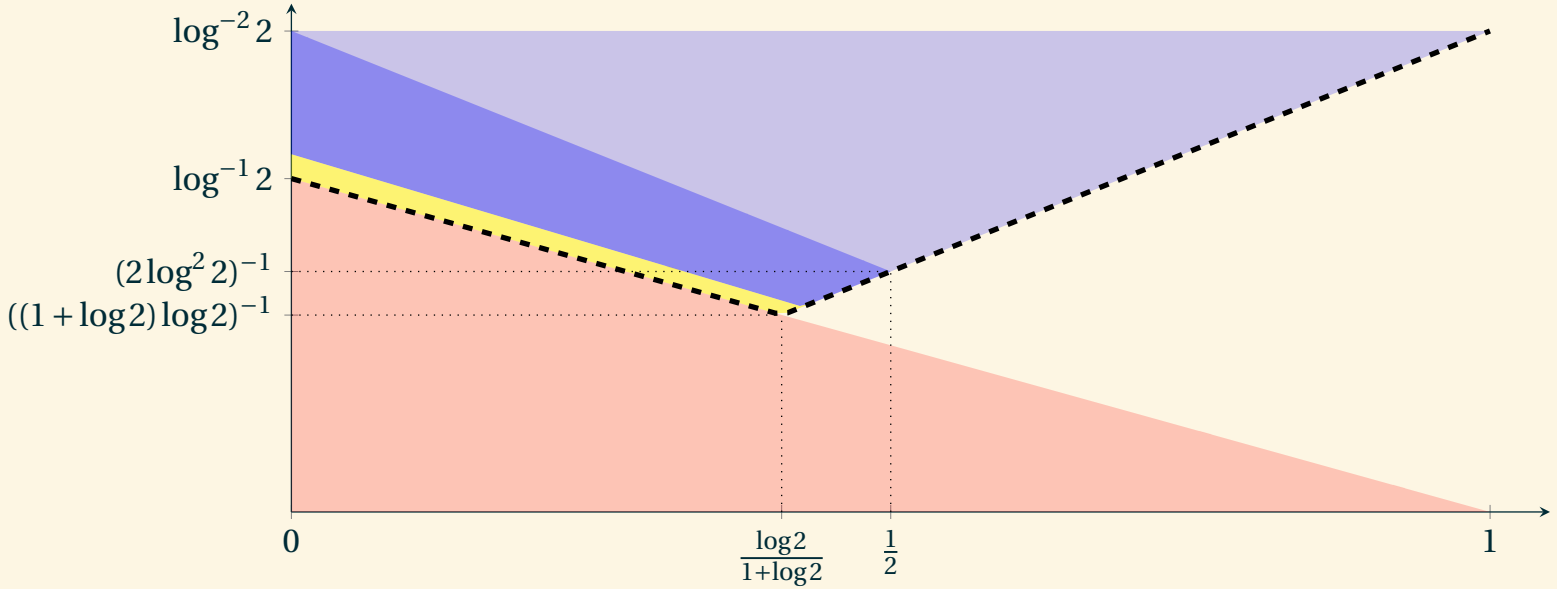
The SPIV algorithm



Unexplained tests

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

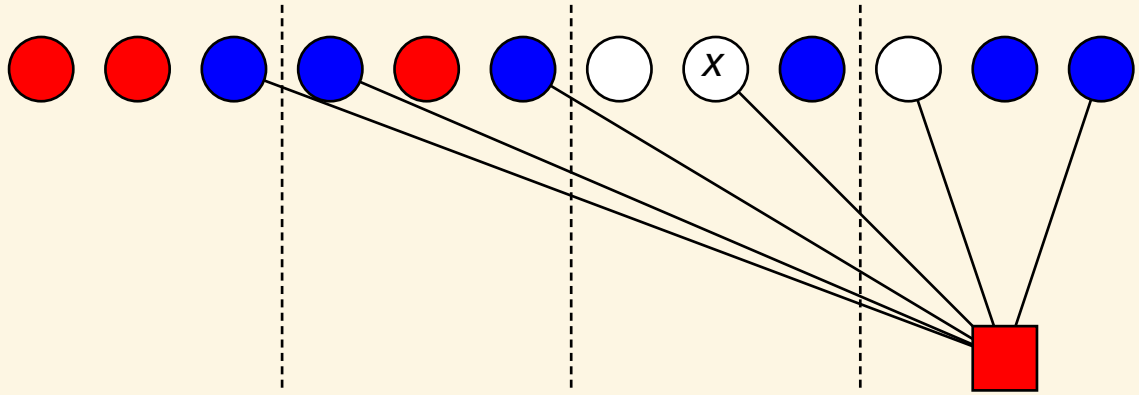
The SPIV algorithm



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

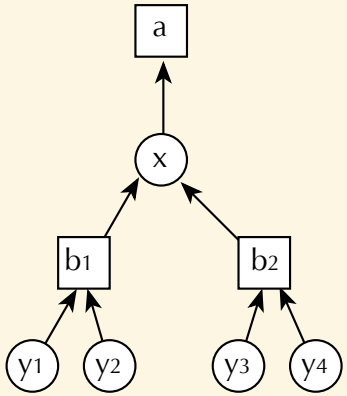
The SPIV algorithm



The score: second attempt

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

The SPIV algorithm



The score: Belief Propagation

$$\mu_{x\to a}(0) \propto (n-k) \prod_{b\in\partial x\setminus a} \mu_{b\to x}(0)$$

$$\mu_{x\to a}(1) \propto k \prod_{b\in\partial x\setminus a} \mu_{b\to x}(1)$$

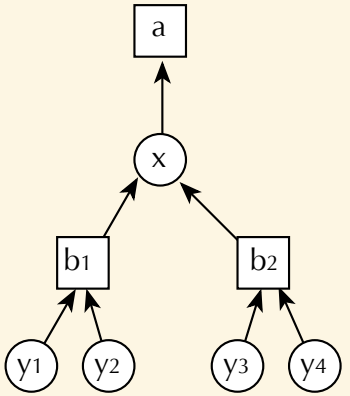
$$\mu_{a\to x}(0\mid 1) \propto 1 - \prod_{y\in\partial a\setminus x} \mu_{y\to a}(0)$$

$$\mu_{a\to x}(1\mid 1) \propto 1$$

$$\mu_{a\to x}(0\mid 0) = 1$$

$$\mu_{a\to x}(1\mid 0) = 0$$

The SPIV algorithm

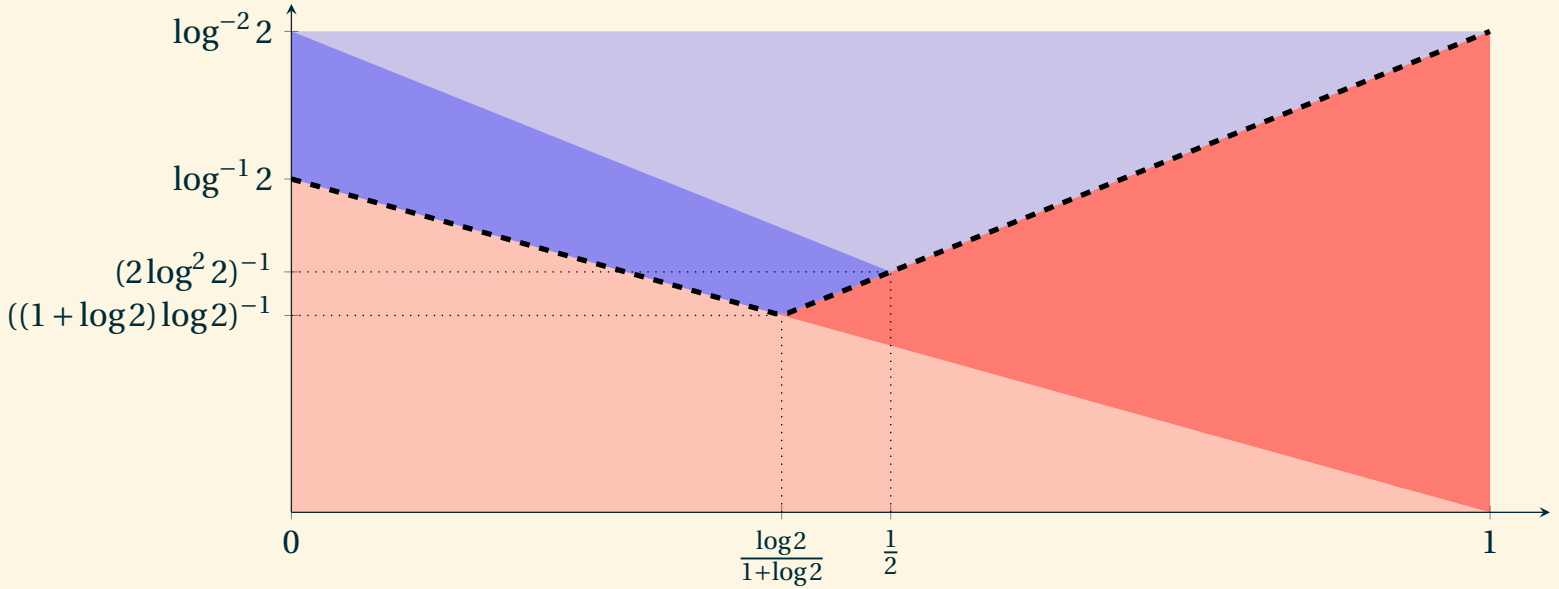


The score: Belief Propagation

$$\eta_{x\to a} = \log\left(\frac{n-k}{k}\right) + \sum_{b\in\partial x\setminus a} \eta_{b\to x}$$

$$\eta_{a\to x} = \log\left(1 - \prod_{y\in\partial a\setminus x} \frac{1+\tanh(\eta_{y\to a}/2)}{2}\right)$$

A matching lower bound



Theorem [COGHKL19]

Identifying the infected individuals is information-theoretically impossible with $(1-\varepsilon)m_{\mathrm{rnd}}$ tests.

A matching lower bound

Proposition [dilution]

Let

$$\frac{\log 2}{1+\log 2} < \theta < \theta' < 1.$$

If there exists a sequence of successful designs for density θ, then there also exists one for θ'.

Proof idea

Add healthy dummies.

A matching lower bound

Proposition

For any $\varepsilon > 0$ there exists $\theta_0(\varepsilon) < 1$ such that for all $\theta_0 < \theta < 1$ and large enough n for any test design with

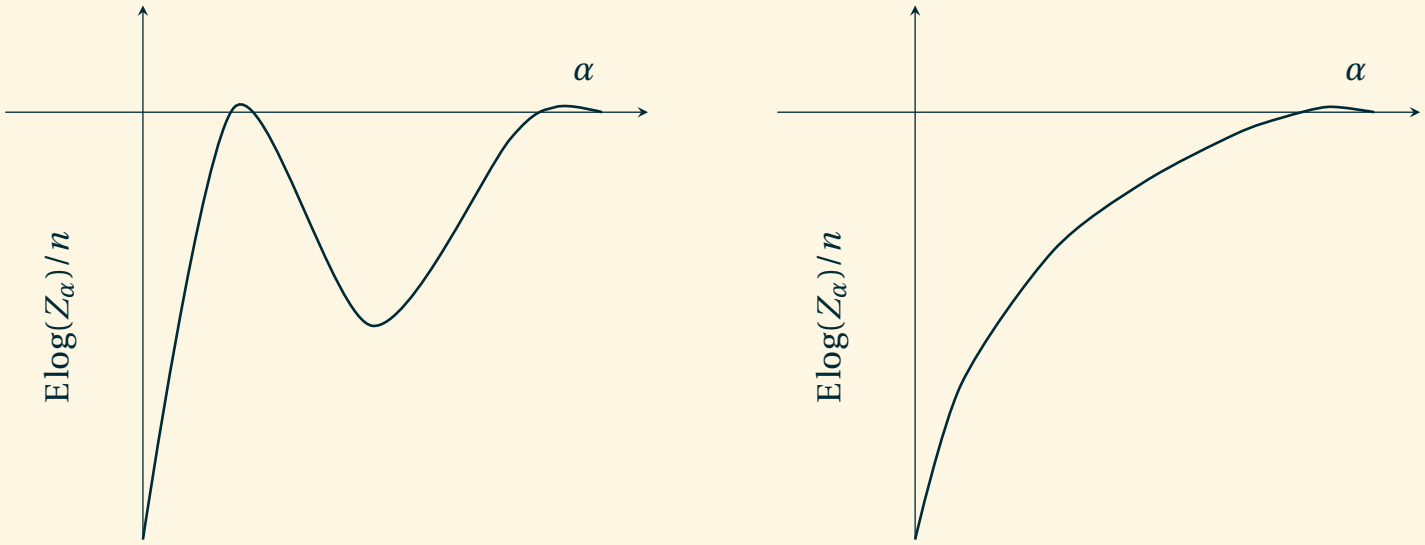
$$m < \frac{\theta - \varepsilon}{\log^2 2} n^\theta \log n$$

tests there are at least $\log n$ disguised individuals w.h.p.

Proof idea

- *Regularisation:* optimal designs are approximately regular
- *Positive correlation:* probability of being disguised [MT11,A18]
- *Probabilistic method:* disguised individuals likely exist

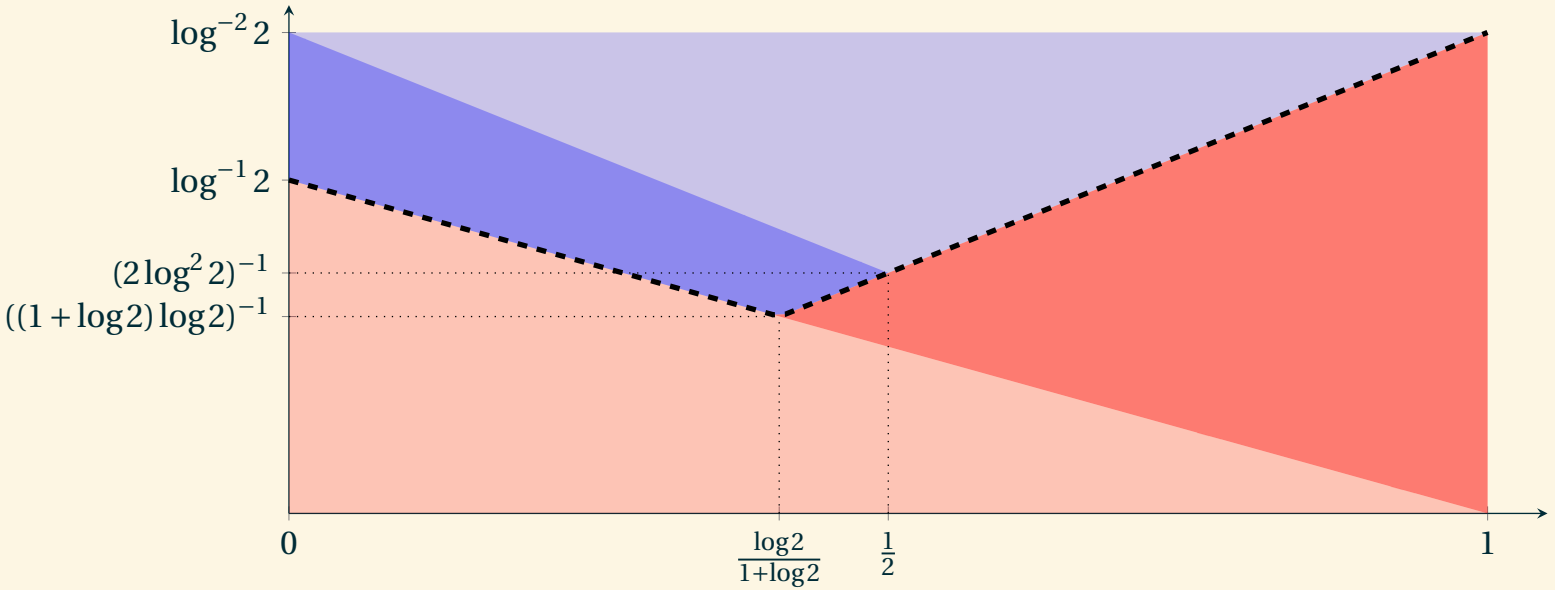
Is spatial coupling necessary?



No overlap gap property vs trivial BP fixed point

- overlap gap in some inference problems [GZ17,BWZ20]
- but not in group testing [IZ20]
- yet BP stuck in trivial fixed point

Summary



- optimal efficient algorithm `SPIV` based on spatial coupling
- matching information-theoretic lower bound
- existence of an adaptivity gap
- optimal two-round adaptive algorithm [HKL19]