Near-Optimal Sparse Adaptive Group Testing

ISIT 2020

Nelvin Tan and Jonathan Scarlett

National University of Singapore (NUS)

June 2020

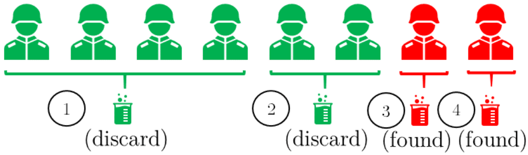
Motivation

- In 1943, the US army had the task of identifying syphilitic soldiers
- Individual blood tests for syphilis were expensive
- Using fewer tests is desirable

Motivation

- Robert Dorfman's key insight: reduce number of tests by pooling

Example:



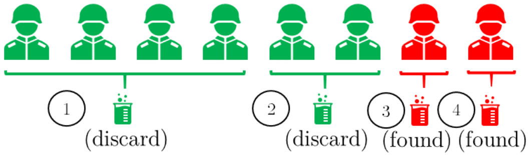
Central problem:

- How many tests are required to accurately discover the infected soldiers?
- How can it be achieved?

Motivation

- Robert Dorfman's key insight: reduce number of tests by pooling

Example:



Central problem:

- How many tests are required to accurately discover the infected soldiers?
- How can it be achieved?

Applications

- **Medical testing:** COVID-19, by pooling Ribonucleic acid (RNA) samples [Yelin et al., 2020]

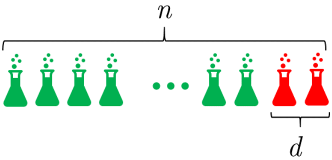
- **Some other applications:**
  - Biology
  - Communications
  - Data science

Applications

- **Medical testing:** COVID-19, by pooling Ribonucleic acid (RNA) samples [Yelin et al., 2020]

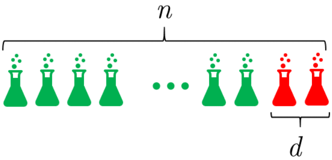
- **Some other applications:**
  - Biology
  - Communications
  - Data science

Group Testing Setup



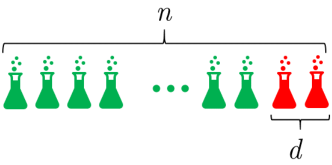
- In this talk:
  - n items labelled $\{1, \ldots, n\}$ that produces binary outcomes when tested
  - Defective set $\mathcal{D} \subset \{1, \ldots, n\}$, where $d = |\mathcal{D}| \in o(n)$
  - Combinatorial prior: Defective set $\mathcal{D} \sim \text{Uniform}\binom{n}{d}$
  - Noiseless testing: negative outcome $\Rightarrow$ all items in pool are non-defective; positive outcome $\Rightarrow$ at least one item in pool is defective
  - Distinction between adaptive and non-adaptive testing

Group Testing Setup



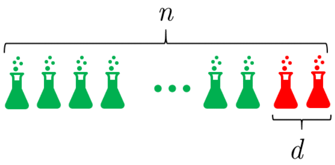
- In this talk:
  - n items labelled $\{1, \dots, n\}$ that produces binary outcomes when tested
  - Defective set $\mathcal{D} \subset \{1, \dots, n\}$, where $d = |\mathcal{D}| \in o(n)$
  - Combinatorial prior: Defective set $\mathcal{D} \sim \text{Uniform}\binom{n}{d}$
  - Noiseless testing: negative outcome $\Rightarrow$ all items in pool are non-defective; positive outcome $\Rightarrow$ at least one item in pool is defective
  - Distinction between adaptive and non-adaptive testing

Group Testing Setup



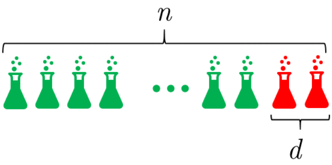
- In this talk:
  - n items labelled $\{1, \ldots, n\}$ that produces binary outcomes when tested
  - Defective set $\mathcal{D} \subset \{1, \ldots, n\}$, where $d = |\mathcal{D}| \in o(n)$
  - Combinatorial prior: Defective set $\mathcal{D} \sim \text{Uniform}\binom{n}{d}$
  - Noiseless testing: negative outcome $\Rightarrow$ all items in pool are non-defective; positive outcome $\Rightarrow$ at least one item in pool is defective
  - Distinction between adaptive and non-adaptive testing

Group Testing Setup



- In this talk:
  - n items labelled $\{1,\dots,n\}$ that produces binary outcomes when tested
  - Defective set $\mathcal{D} \subset \{1,\dots,n\}$, where $d = |\mathcal{D}| \in o(n)$
  - Combinatorial prior: Defective set $\mathcal{D} \sim \text{Uniform}\binom{n}{d}$
  - Noiseless testing: negative outcome $\Rightarrow$ all items in pool are non-defective; positive outcome $\Rightarrow$ at least one item in pool is defective
  - Distinction between adaptive and non-adaptive testing

Group Testing Setup



- In this talk:
  - n items labelled $\{1, \dots, n\}$ that produces binary outcomes when tested
  - Defective set $\mathcal{D} \subset \{1, \dots, n\}$, where $d = |\mathcal{D}| \in o(n)$
  - Combinatorial prior: Defective set $\mathcal{D} \sim \text{Uniform}\binom{n}{d}$
  - Noiseless testing: negative outcome $\Rightarrow$ all items in pool are non-defective; positive outcome $\Rightarrow$ at least one item in pool is defective
  - Distinction between adaptive and non-adaptive testing

Recovery Criteria

- Error probability bounded by some $\epsilon > 0$:

$$P_e := \mathbb{P}[\widehat{\mathcal{D}} \neq \mathcal{D}] \leq \epsilon$$

- We study two conditions on the number of tests T:
  - Information theoretic lower bound:
    - Necessary number of tests T for $P_e \leq \epsilon$
  - Upper bound from algorithm:
    - Sufficient number of tests T our algorithm needs for $P_e \leq \epsilon$

Recovery Criteria

- Error probability bounded by some $\epsilon > 0$:

$$P_e := \mathbb{P}[\widehat{\mathcal{D}} \neq \mathcal{D}] \leq \epsilon$$

- We study two conditions on the number of tests T:
  - Information theoretic lower bound:
    - Necessary number of tests T for $P_e \leq \epsilon$
  - Upper bound from algorithm:
    - Sufficient number of tests T our algorithm needs for $P_e \leq \epsilon$

Sparse Group Testing

Testing procedure is subjected to one of the following:

- Items are finitely divisible and thus may participate in at most γ tests
- Tests are size-constrained and thus contain no more than ρ items per test

Example: We need at least 20 *ml* of blood per soldier for reliable testing.

- Divisibility constraint: finite amount of blood per soldier
- Size constraint: limitations on volume capacity of the machine

Sparse Group Testing

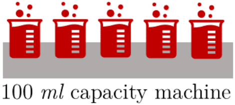
Testing procedure is subjected to one of the following:

- Items are finitely divisible and thus may participate in at most γ tests
- Tests are size-constrained and thus contain no more than ρ items per test

Example: We need at least 20 *ml* of blood per soldier for reliable testing.

- Divisibility constraint: finite amount of blood per soldier

- Size constraint: limitations on volume capacity of the machine



Sparse Group Testing

- Previous work in literature shows that $\gamma \in \Theta(\log n)$ and $\rho \in \Theta\left(\frac{n}{d}\right)$ are required to attain optimal scaling laws for the unconstrained setting
- We are interested in:
  - Divisibility constraint: $\gamma \in o(\log n)$
  - Size constraint: $\rho \in o\left(\frac{n}{d}\right)$

Sparse Group Testing

- Previous work in literature shows that $\gamma \in \Theta(\log n)$ and $\rho \in \Theta\left(\frac{n}{d}\right)$ are required to attain optimal scaling laws for the unconstrained setting
- We are interested in:
  - Divisibility constraint: $\gamma \in o(\log n)$
  - Size constraint: $\rho \in o\left(\frac{n}{d}\right)$

Previous Work on Non-Adaptive Setting

- Considered the non-adaptive setting
- For some error probability $P_e = \mathbb{P}[\widehat{\mathcal{D}} \neq \mathcal{D}] \leq \epsilon$:

Constraint type	Scaling regime	Tests required
γ-divisible items	$d \in \Theta(n^\theta), \theta < 1$	$T > \gamma d \left(\frac{n}{d}\right)^{(1-5\epsilon)/\gamma}$
	$\gamma \in o(\log n)$	$T < \left\lceil \gamma d \left(\frac{n}{\epsilon}\right)^{1/\gamma} \right\rceil$
ρ-sized tests	$d \in \Theta(n^\theta), \theta < 1$	$T \in \Omega\left(\frac{n}{\rho}\right)$
	$\rho \in o\left(\frac{n}{d}\right)$	$T \in O\left(\frac{n}{\rho}\right)$

Table: Previous results (non-adaptive setting)

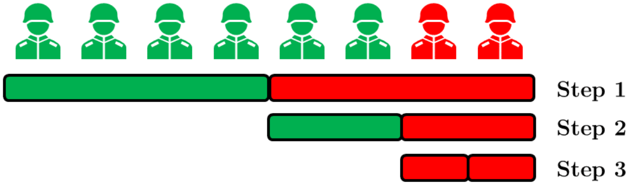
Outline

1. Adaptive Setting for γ-Divisible Items
   - Lower Bound Result
   - Upper Bound Result
2. Overview of Results

Adaptive Setting

Adaptive setting: test pools are designed sequentially, and each one can depend on previous test outcomes.

Example:



Lower Bound Result

Theorem

If $d \in o(n)$, $\gamma \in o(\log n)$, any non-adaptive or adaptive group testing algorithm that tests each item at most γ times and has $P_e \leq \epsilon$ requires at least $e^{-(1+o(1))}\gamma d\left(\frac{n}{d}\right)^{1/\gamma}$ tests.

Improvements: We have strengthened previous lower bound by

- Improving dependence on ϵ, and
- Extending its validity to the adaptive setting.

Lower Bound Result

Theorem

If $d \in o(n)$, $\gamma \in o(\log n)$, any non-adaptive or adaptive group testing algorithm that tests each item at most γ times and has $P_e \leq \epsilon$ requires at least $e^{-(1+o(1))}\gamma d\left(\frac{n}{d}\right)^{1/\gamma}$ tests.

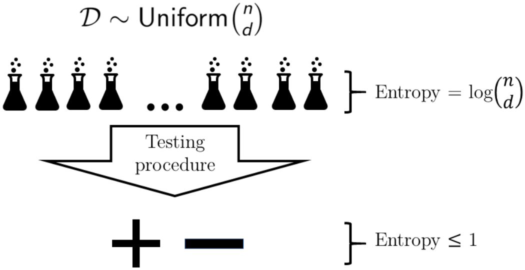
Improvements: We have strengthened previous lower bound by

- Improving dependence on ϵ, and
- Extending its validity to the adaptive setting.

Lower Bound Result Interpretation

Theorem: We require at least $e^{-(1+o(1))\gamma} d \left(\frac{n}{d}\right)^{1/\gamma}$ tests.

Interpretation:

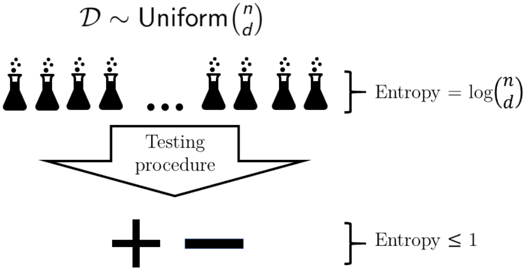


- If every test reveals 1 bit of entropy, we need $\log\binom{n}{d} \approx d \log\left(\frac{n}{d}\right)$ tests
- Our constraint results in tests to be less informative $\Rightarrow$ need more tests than unconstrained setting

Lower Bound Result Interpretation

Theorem: We require at least $e^{-(1+o(1))}\gamma d\left(\frac{n}{d}\right)^{1/\gamma}$ tests.

Interpretation:



- If every test reveals 1 bit of entropy, we need $\log\binom{n}{d} \approx d\log\left(\frac{n}{d}\right)$ tests
- Our constraint results in tests to be less informative $\Rightarrow$ need more tests than unconstrained setting

Lower Bound Proof Outline

- Using a counting argument, we get

$$\mathbb{P}[\text{suc}] \leq \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}},$$

where intuitively,

  - numerator: # possible test outcomes
  - denominator: # defective sets of size d

- From an asymptotic analysis of the counting-based bound, we obtain our lower bound for T

Lower Bound Proof Outline

- Using a counting argument, we get

$$\mathbb{P}[\text{suc}] \leq \frac{\sum_{i=0}^{\gamma d} \binom{T}{i}}{\binom{n}{d}},$$

  where intuitively,
  - numerator: # possible test outcomes
  - denominator: # defective sets of size d
- From an asymptotic analysis of the counting-based bound, we obtain our lower bound for T

Upper Bound Result

Theorem

If $d \in o(n)$, $\gamma \in o(\log n)$, then there exists an adaptive group testing algorithm that tests each item at most γ times achieving $P_e = 0$ using at most $T = \gamma d (\frac{n}{d})^{1/\gamma}$ tests.

Improvements:

- Improved scaling over previous non-adaptive result: $\left(\frac{n}{\epsilon}\right)^{1/\gamma} \Rightarrow \left(\frac{n}{d}\right)^{1/\gamma}$
- Matches the lower bound $T \geq e^{-(1+o(1))}\gamma d\left(\frac{n}{d}\right)^{1/\gamma}$ up to a constant factor of $e^{-(1+o(1))}$

Upper Bound Result

Theorem

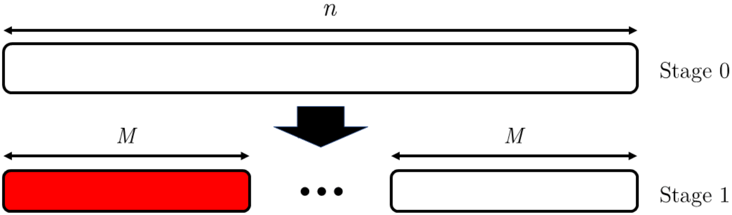
If $d \in o(n)$, $\gamma \in o(\log n)$, then there exists an adaptive group testing algorithm that tests each item at most γ times achieving $P_e = 0$ using at most $T = \gamma d(\frac{n}{d})^{1/\gamma}$ tests.

Improvements:

- Improved scaling over previous non-adaptive result: $\left(\frac{n}{\epsilon}\right)^{1/\gamma} \Rightarrow \left(\frac{n}{d}\right)^{1/\gamma}$
- Matches the lower bound $T \geq e^{-(1+o(1))}\gamma d\left(\frac{n}{d}\right)^{1/\gamma}$ up to a constant factor of $e^{-(1+o(1))}$

Adaptive Algorithm

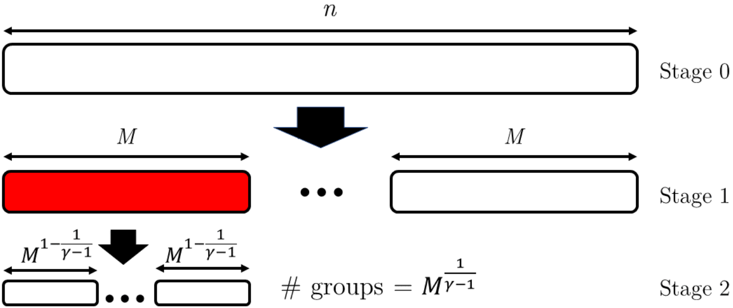
Key idea: Can we partition the items into equal groups of ideal sizes?



- **Group sizes:** $M \to M^{1-\frac{1}{\gamma-1}} \to M^{1-\frac{2}{\gamma-1}} \to \cdots \to 1$
  - n/M splits from stage 0 to stage 1
  - $M^{1/(\gamma-1)}$ splits between any two subsequent stages

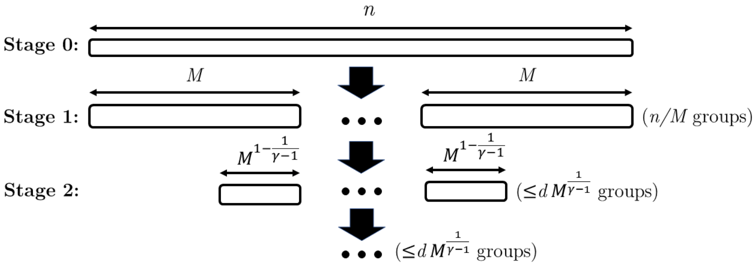
Adaptive Algorithm

Key idea: Can we partition the items into equal groups of ideal sizes?



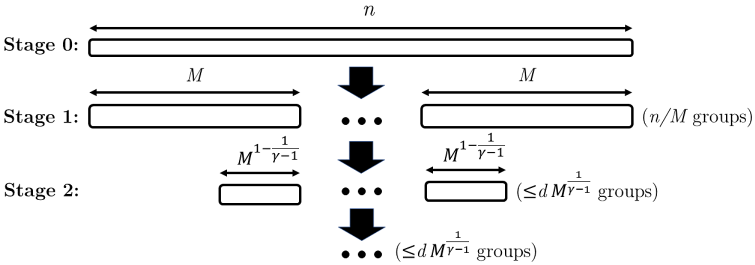
- **Group sizes:** $M \to M^{1-\frac{1}{\gamma-1}} \to M^{1-\frac{2}{\gamma-1}} \to \cdots \to 1$
  - n/M splits from stage 0 to stage 1
  - $M^{1/(\gamma-1)}$ splits between any two subsequent stages

Adaptive Algorithm Analysis



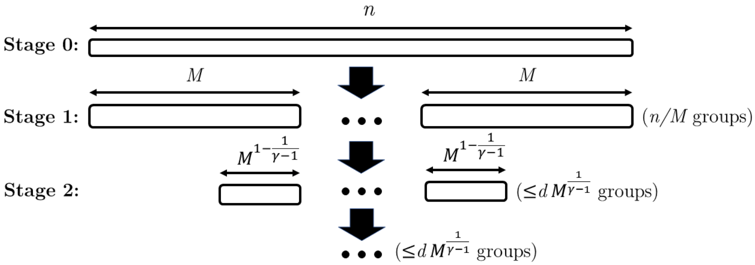
- In stage 1: we made n/M tests
- From stage 2 onwards: we made at most $dM^{\frac{1}{\gamma-1}}$ tests
- This gives us $T \leq \frac{n}{M} + (\gamma-1)dM^{\frac{1}{\gamma-1}}$.
- Optimizing the upper bound w.r.t. M gives us $M = \left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}$
- Substituting back into the upper bound, we get our result

Adaptive Algorithm Analysis



- In stage 1: we made n/M tests
- From stage 2 onwards: we made at most $dM^{\frac{1}{\gamma-1}}$ tests
- This gives us $T \le \frac{n}{M} + (\gamma - 1)dM^{\frac{1}{\gamma-1}}$.
- Optimizing the upper bound w.r.t. M gives us $M = \left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}$
- Substituting back into the upper bound, we get our result

Adaptive Algorithm Analysis



- In stage 1: we made n/M tests
- From stage 2 onwards: we made at most $dM^{\frac{1}{\gamma-1}}$ tests
- This gives us $T \leq \frac{n}{M} + (\gamma - 1)dM^{\frac{1}{\gamma-1}}$.
- Optimizing the upper bound w.r.t. M gives us $M = \left(\frac{n}{d}\right)^{\frac{\gamma-1}{\gamma}}$
- Substituting back into the upper bound, we get our result

Overview of Results

Recap: Focused on the adaptive setting with γ-divisible items constraint.

	Scaling regime	Tests required
non-adaptive	$d \in \Theta(n^\theta), \theta < 1$ $\gamma \in o(\log n)$	$T > \gamma d \left(\frac{n}{d}\right)^{(1-5\epsilon)/\gamma}$ $T < \left\lceil \gamma d \left(\frac{n}{\epsilon}\right)^{1/\gamma} \right\rceil$
adaptive	$d \in o(n)$ $\gamma \in o(\log n)$ $\gamma d \to \infty$	$T > e^{-(1+o(1))} \gamma d \left(\frac{n}{d}\right)^{1/\gamma}$ $T < \gamma d \left(\frac{n}{d}\right)^{1/\gamma}$

Table: Comparison of our adaptive results with previous non-adaptive results

Overview of Results

Recap: Focused on the adaptive setting with γ-divisible items constraint.

	Scaling regime	Tests required
non-adaptive	$d \in \Theta(n^\theta), \theta < 1$ $\gamma \in o(\log n)$	$T > \gamma d \left(\frac{n}{d}\right)^{(1-5\epsilon)/\gamma}$ $T < \left\lceil \gamma d \left(\frac{n}{\epsilon}\right)^{1/\gamma} \right\rceil$
adaptive	$d \in o(n)$ $\gamma \in o(\log n)$ $\gamma d \to \infty$	$T > e^{-(1+o(1))} \gamma d \left(\frac{n}{d}\right)^{1/\gamma}$ $T < \gamma d \left(\frac{n}{d}\right)^{1/\gamma}$

Table: Comparison of our adaptive results with previous non-adaptive results