Group Testing with Side Information

Dror Baron[n]

with Junan Zhu,[h] Kristina Rivera,[n] Shujie Cao,[w] Chau-Wai Wong,[n] Ritesh Goenka,[i] and Ajit Rajwade[i]

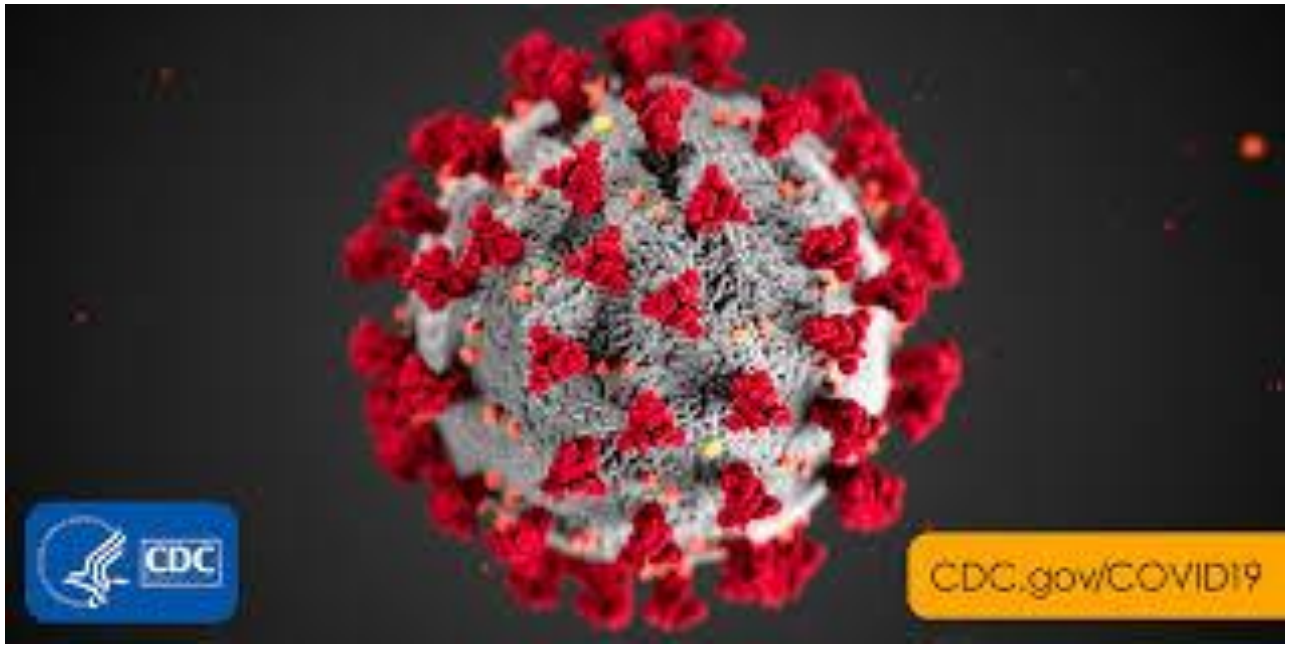
[h]Harvest Fund, [i]IIT Bombay, [n]NC State University, and [w]Northwestern University

April 2023






Motivation



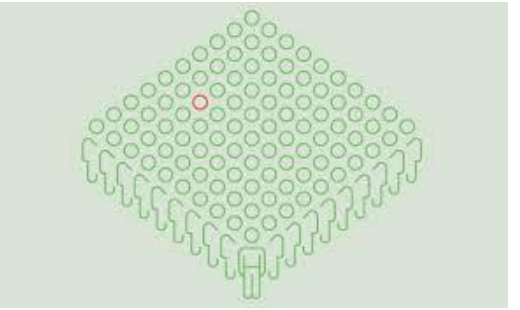
- Fast / efficient / affordable testing of large populations

Conventional testing

- Swab patient for mucus
  - Saliva also feasible [Wyllie et al. 2020]
- Amplify viral material
  - Can use reverse transcription polymerase chain reaction (RT-PCR)
- Test for viral material

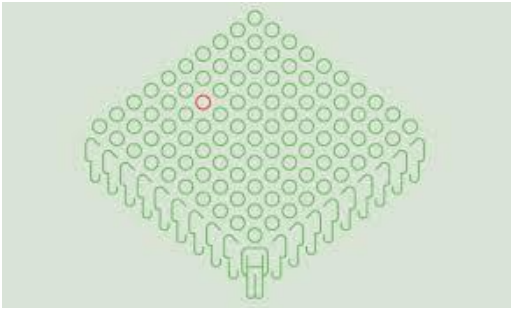
- Challenges
  - False negatives / positives
  - Time / resource intense
  - How much testing? [Kontoyiannis et al. 2020]
- Want fast / efficient / affordable testing

Pooled / group testing [Dorfman 1943]



- Suppose low prevalence (0.1%? 1%?); few people sick
- Pool group of (10?) people's samples together
- All healthy → negative pooled test → rules out group
- Any sick → positive → need more information

Pooled / group testing [Dorfman 1943]



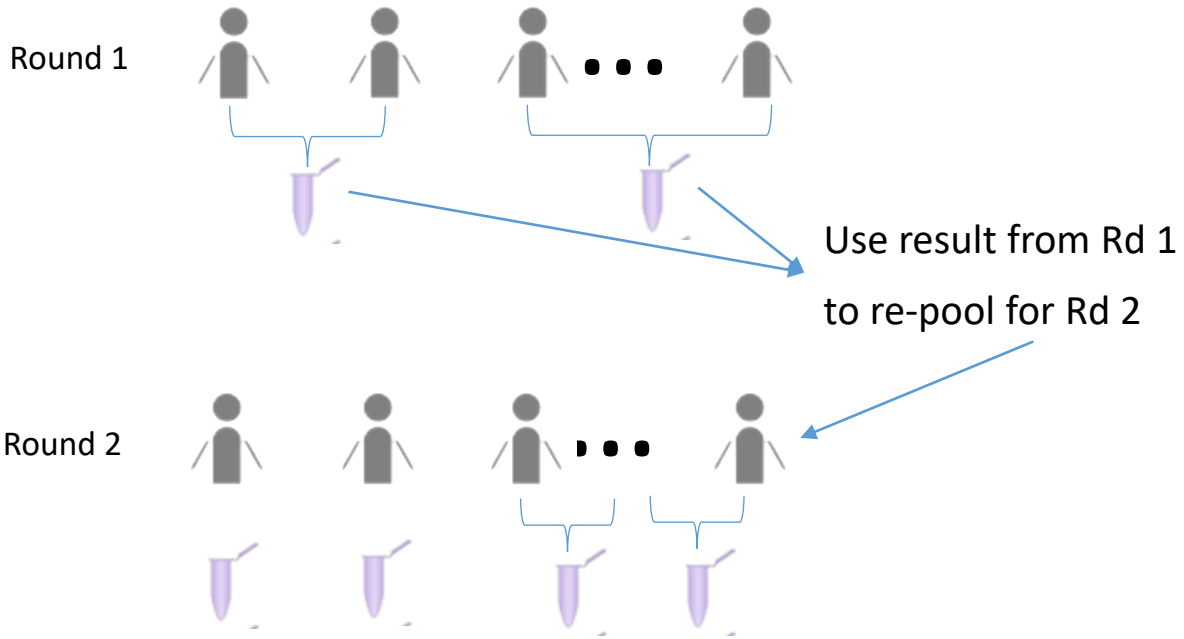
- Suppose low prevalence (0.1%? 1%?); few people sick
- Pool group of (10?) people's samples together
- All healthy → negative pooled test → rules out group
- Any sick → positive → need more information

- Can optimize pool size [Hanel & Thurner 2020]
- Demonstrated for COVID-19 [Kishony et al. 2020]
  - Used in Nebraska [Bilder]; China tested millions daily in 2022
- Testing frequency vs. disease spread risk [Lakdawalla et al. 2020]
  - Test asymptomatics (no symptoms) frequently
- Non-adaptive approaches (Tapestry; IIT Bombay; [Ghosh et al. 2020])

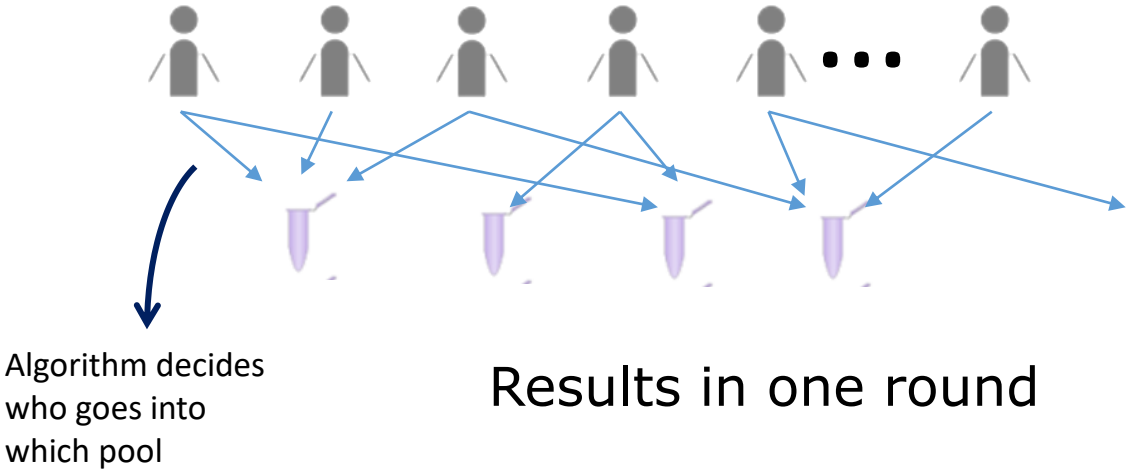
Non-Adaptive Group Testing

Non-adaptive group testing

- Dorfman pooling
- Each sample in *single pool*



- *Each sample in multiple pools*
- Identify positive individuals by combining tests
- Single-round (non-adaptive)



Results in one round

Robust to erroneous tests

- Dorfman
  - False negative in first round → doesn't reach second round
  - *Very sensitive to erroneous tests / dilution / etc.*

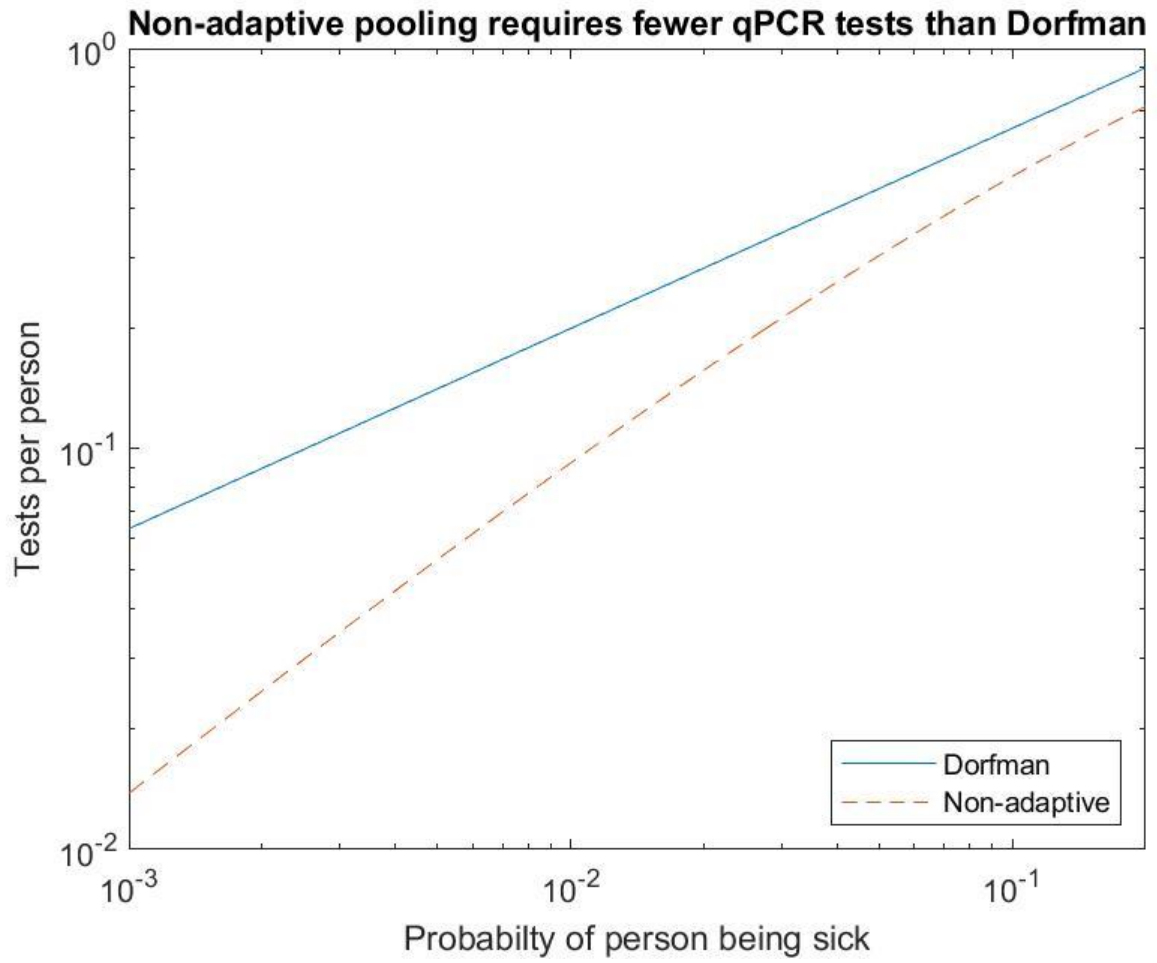
- Non-adaptive pooling
  - Suppose (example) each individual sample goes to 5 pools
  - All 5 positive → individual very likely positive
  - 0-3 positive → very likely negative
  - 4? Depends on Probability(false negative) & pool structure
  - Algorithm fuses information into probabilities

- *Robust to erroneous tests → dilution less important*

Dorman versus non-adaptive

- Dorfman sensitive to errors → use small pools → more tests
- Non-adaptive uses larger pools → fewer PCR tests
  - Pool sizes up to 48

- *Big edge at low prevalence*
  - Great for asymptomatics
  - Also better at high prevalence



Lower latency

- Dorfman
  - Moderate test capacity improvement → wait hours for PCR machine
  - Needs 2 rounds (e.g., 3-hour PCR → 6 hours)

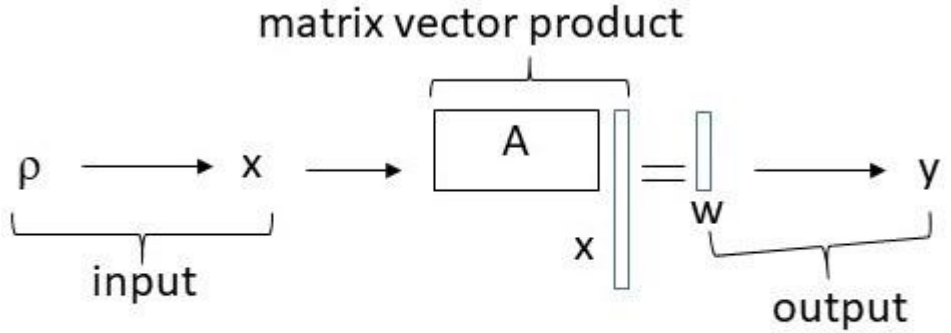
- Non-adaptive pooling
  - Big capacity improvement → PCR machines immediately available
  - Non-adaptive techniques use *single round* (3 hours)
- *Lower latency overall*

- Can use "semi-adaptive" multi-round pooling
  - More latency but fewer misdiagnoses

How?

Problem Formulation and Measurement Channel

Problem formulation



- Convert to linear algebra

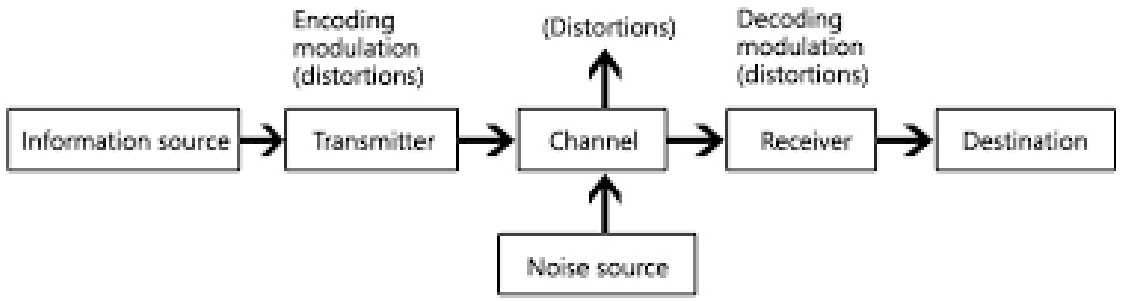
- ρ sickness prevalence
- x *input vector*
  - Binary (x_n=1 sick; x_n=0 healthy) or real valued viral loads

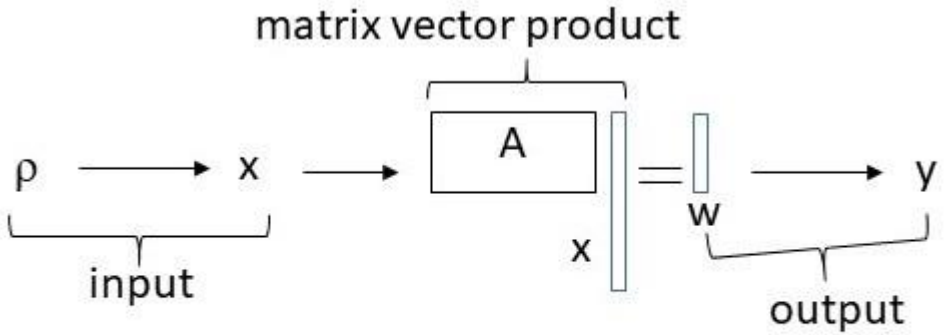
- Multiply x by binary *measurement matrix* A
  - Rows/cols correspond to measurements / patients

- Matrix vector product w; w_m #sick in measurement m
- *Noisy* y_m depends on w_m

- Goal: Estimate x from y, A, statistical info (e.g., ρ)

Communication system analogy





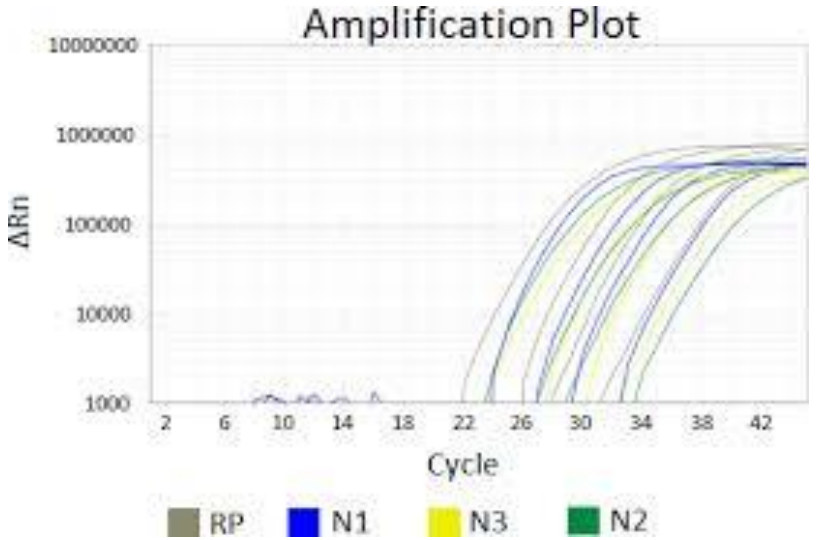
- Want to communicate x
- Encoder converts x to w
- Transmit w over noisy channel
- Noisy channel output y
- Recover x from y

- x patient status vector
- Auxiliary vector w=Ax
- Measurement y_m depends on w_m
- Measurement channel $f(y_m|w_m)$
- Recover x from y, A, ρ, $f(y_m|w_m)$

- *Want encoder A and decoding algo to maximize information flow from x to y*

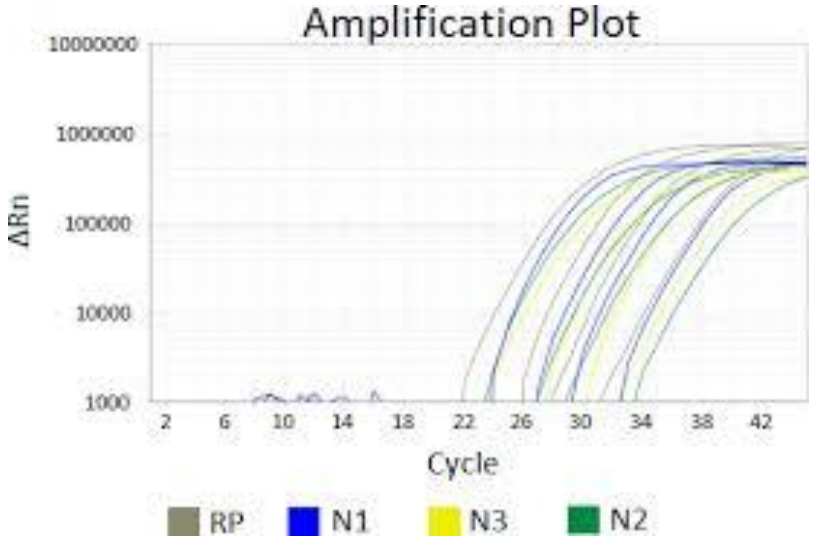
What channel?

- More about RT-PCR
  - Genetic test
  - Viral density increase ~2X per iteration
  - Sufficiently large viral density → fluorescent



- When is it fluorescent? (Tapestry; IIT Bombay; [Ghosh et al. 2020])
  - No viral matter → never
  - Minimal → 37-38 iterations
  - Sick patient → 22-31 iterations

Two PCR channels



- *Binary* PCR
  - Several dozen iterations
  - Fluorescent (1) or not (0)?
  - Binary input & output
  - False positive – contamination is amplified
  - False negative – weak viral load diluted by pooling

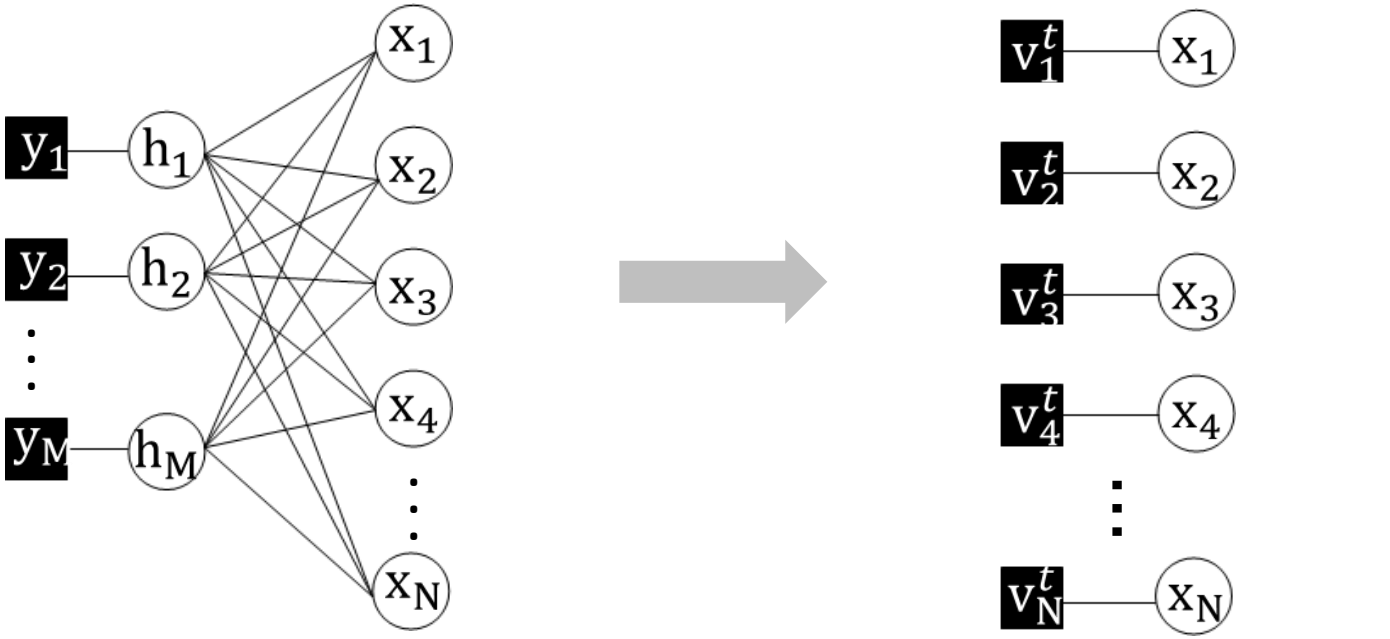
- *Quantitative* PCR (Tapestry; IIT Bombay; [Ghosh et al. 2020])
  - Multiplicative noise: $\log_2(y_m)=\log_2(w_m)+N(0,0.01)$
    - Special case, $w_m=0 \rightarrow y_m=0$
  - Non-negative real input & output

Approximate Message Passing

Linear regression for large “well-behaved” matrices

Approximate message passing [Donoho et al. 2009]

- Fast iterative algorithm
- Decouples matrix problem (y=w+z=Ax+z, Gaussian z) to simpler scalar channel denoising (v=x+Gaussian noise)
- Based on approximation of precise message passing



$$y = h + z = Ax + z$$

$$v = x + \mathit{noise}$$

AMP steps

- Initialize $x^0=0$
- At iteration t, do
- Residual: $r^t = y - Ax^t + \frac{r^{t-1}}{M/N}\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1})\rangle$
- Pseudo-data: $v^t = x^t + A^T r^t$
- Denoising: $x^{t+1} = \eta_t(v^t)$

AMP steps

- Initialize $x^0=0$
- At iteration t, do
- Residual: $r^t = y - Ax^t + \frac{r^{t-1}}{M/N}\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1})\rangle$
- Pseudo-data: $v^t = x^t + A^T r^t$
- Denoising: $x^{t+1} = \eta_t(v^t)$

Denoising function often $\eta = E[X|V]$

AMP steps

- Initialize $x^0=0$
- At iteration t, do
- Residual: $r^t = y - Ax^t + \frac{r^{t-1}}{M/N}\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1})\rangle$

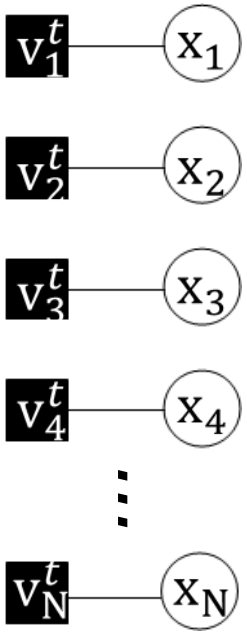
Onsager correction term ensures v=x+Gaussian

- Pseudo-data: $v^t = x^t + A^T r^t$
- Denoising: $x^{t+1} = \eta_t(v^t)$

AMP steps

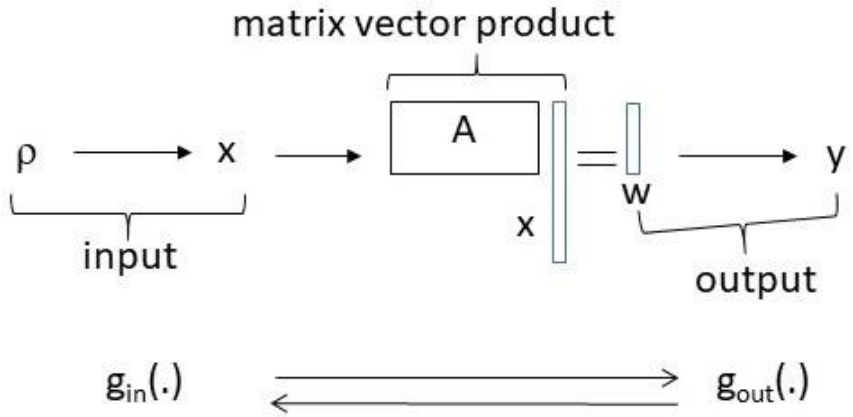
- Initialize $x^0=0$
- At iteration t, do
- Residual: $r^t = y - Ax^t + \frac{r^{t-1}}{M/N}\langle \eta'_{t-1}(x^{t-1} + A^T r^{t-1})\rangle$
- Pseudo-data: $v^t = x^t + A^T r^t$
- Denoising: $x^{t+1} = \eta_t(v^t)$

Standard AMP: $\eta_t(v^t)$ is scalar



Generalized AMP (GAMP)

[Rangan 2011]

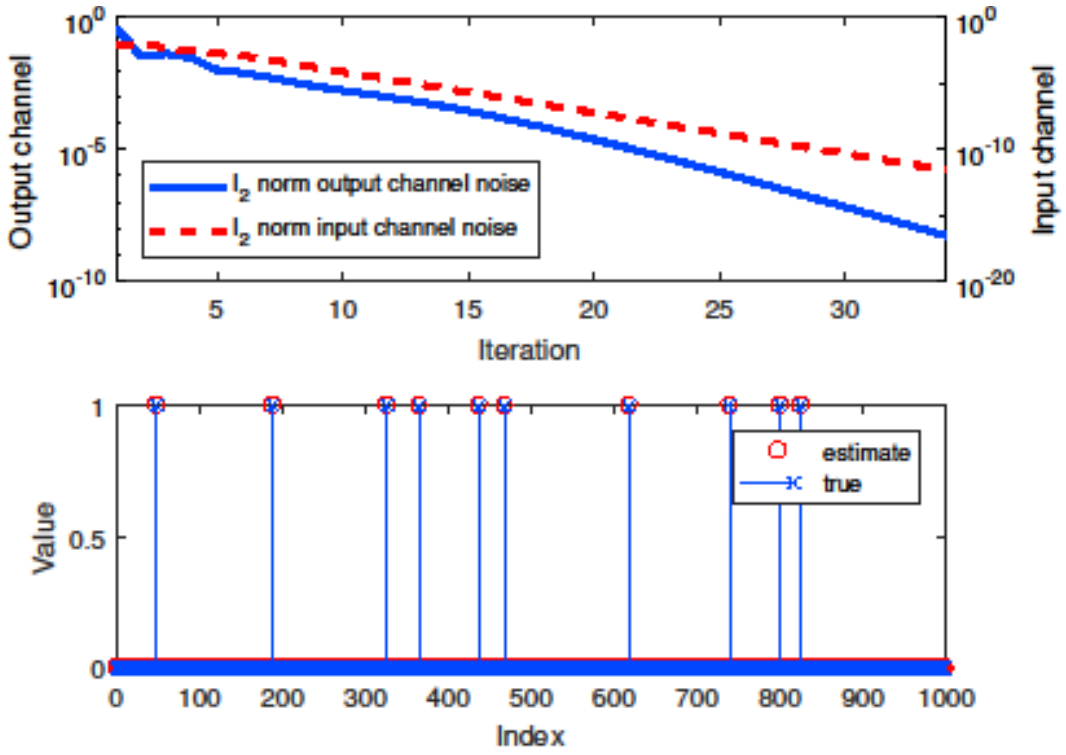


- Recall w=Ax
  - AMP: y=w+Gaussian
  - GAMP: probability density f(y|w)
  - Resembles AMP; also iteratively denoises w
  - Input / output channels
    - $g_{in}(.)$ between prior & x; $g_{out}(.)$ between y & w

- Examples:
  1. Additive noise, y=w+z, Gaussian z; use AMP
  2. Binary PCR → binary y [Zhu, B, Rivera 2020]
  3. Quantitative PCR → real-valued y

Numerical results [Zhu, B, Rivera 2020]

- Binary channel
- N=5000 patients; ρ=1% prevalence; M=1000 measurements
- R=M/N=20% measurement rate



- Accurate reconstruction
- Fast (~1 sec on laptop)

Side Information

Side information

- Earlier goal: Estimate x from y, A, statistical info
- Side information (SI) often available
  - Symptoms affect probability of infection
  - Family members w/ correlated infection status
  - Address, profession, coworkers, …
  - Contact tracing

- Input x no longer independent and identically distributed (iid)
  - *Non-identical* distributions (symptoms, address, …)
  - *Dependencies* between variables (families, contact tracing, …)

GAMP with SI for non-iid x

- AMP can use SI in denoiser

  [B et al. 2017, Ma et al. 2019, Liu et al. 2020, Liu et al. 2022]

- Vector denoisers support dependencies between patients
  [Donoho et al. 2013, Ma et al. 2014]

GAMP with SI for non-iid x

- AMP can use SI in denoiser

  [B et al. 2017, Ma et al. 2019, Liu et al. 2020, Liu et al. 2022]

- Vector denoisers support dependencies between patients

  [Donoho et al. 2013, Ma et al. 2014]

- Contribution to group testing community:
  - Prior art considers non-identical distributions
  - Dependent variables with combinatorial complexity [Cuturi et al. 2020]
  - Group testing for connected communities [Nikolopoulos et al. 2020]
  - Our approach supports various non-i.i.d. distributions and is fast

GAMP with SI for non-iid x

- AMP can use SI in denoiser

  [B et al. 2017, Ma et al. 2019, Liu et al. 2020, Liu et al. 2022]

- Vector denoisers support dependencies between patients

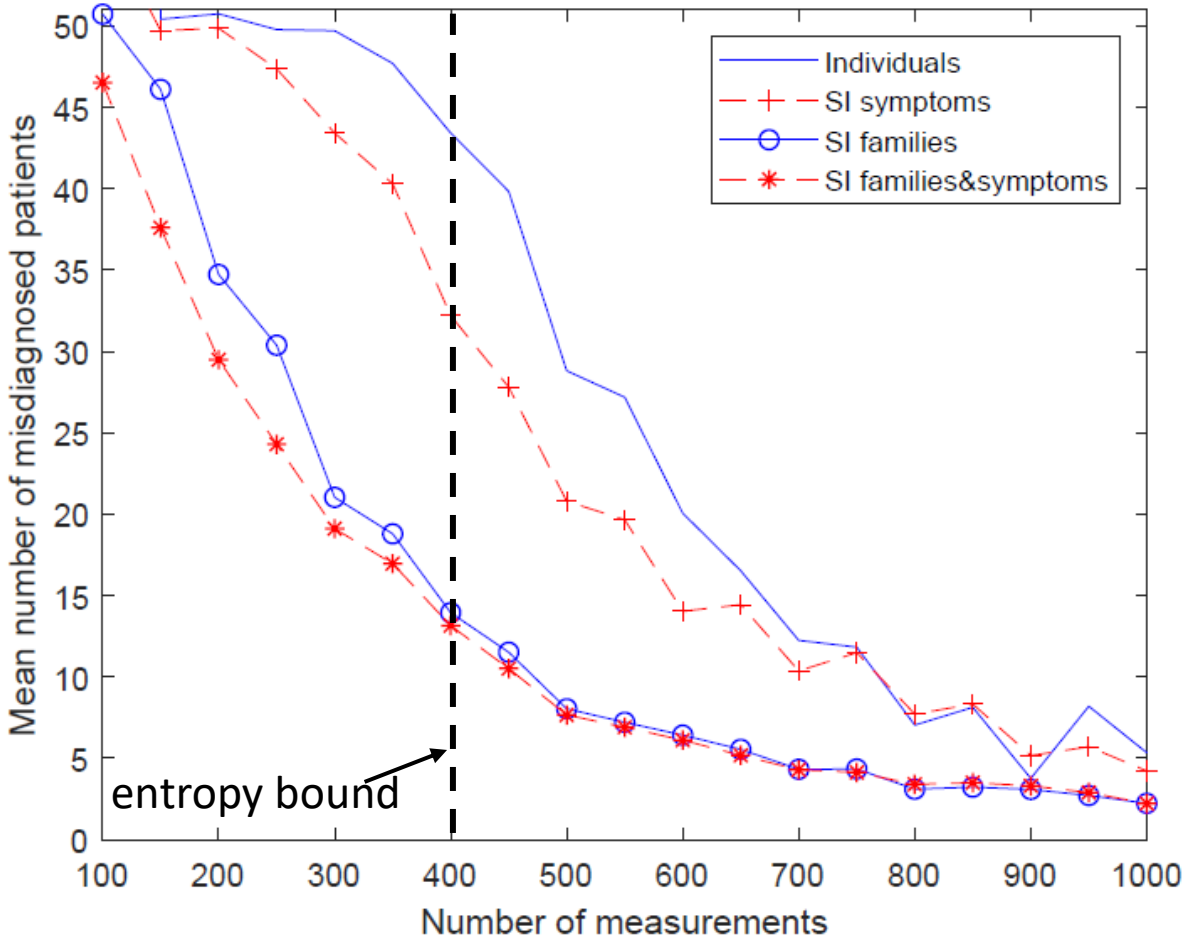
  [Donoho et al. 2013, Ma et al. 2014]

- Contribution to AMP community:
  - Vector denoisers with SI in GAMP
  - Numerical evidence for binary A with const ones per row/col

Numerical Results

Numerical results – family and symptom SI

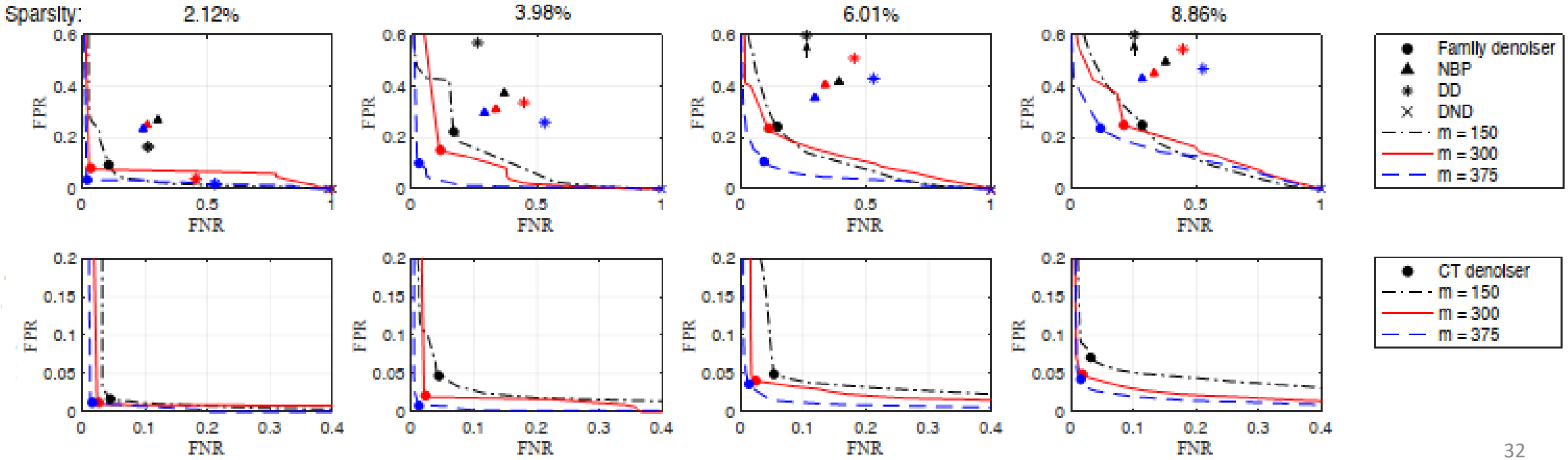
- As before, N=5000 patients; ρ=1% prevalence
- Families SI (family size F=4) *dependencies*
  - ρ_f=1.5% of families infected; ρ_i=2/3 of individuals within families
- Symptoms SI *non-identical*
  - Families w/symptoms ρ_1=5%
  - Families without ρ_2=1%
- *R below entropy bound*
  - SI reduces entropy
- Both types of SI help
- Dependencies more useful



Numerical results – contract tracing SI

[Cao et al. 2022]

- Contact tracing (CT) and infections data simulated using susceptible, exposed, infectious, recovered (SEIR) model
- Compared to nonparametric belief propagation (NBP), definite defectives (DD), definitely nondefective (DND)
- CT SI (row2) > Family SI (row1) > {NBP, DD, DND}



Discussion

- Analogy between communication system and viral testing
- Contributions
  - Convert pooled tests to noisy linear inverse problem
  - GAMP solver
  - Use SI in GAMP; supports dependencies between patients
  - Contact tracing SI – more dependencies further reduce # measurements
- More - matrix design seems to matter less than decoding algo
- Future work
  - GAMP for quantitative PCR (multiplicative noise)

Thanks!

More details in our papers http://barondror.com/

If you liked the video, please like it and subscribe
https://www.youtube.com/channel/UCKy5Pyk8pmxXGu316UU34VQ