Near optimal efficient decoding from sparse pooled data

arXiv:2108.04342

Max Hahn-Klimroth & Noela Müller

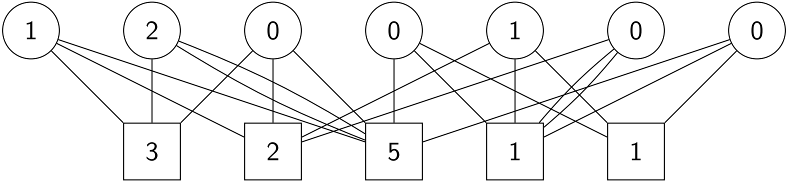
August 12th, 2021

Questions

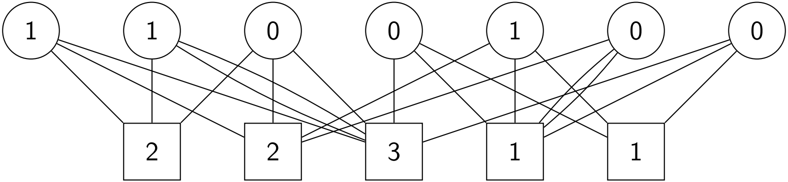
- What is *pooled data*?
- What is *sparse*?
- What is *near optimal* in this context?
- How does it work?

Pooled data

- n items $x_1, \ldots, x_n$, each of a specific weight $\boldsymbol{\sigma}_i \in \{0, 1, \ldots, d\}$.
- A *ground-truth* or *signal* $\boldsymbol{\sigma} \in \{0, 1, \ldots, d\}^n$ is drawn uniformly from a probability distribution.
- We can pool items together and measure them (additive model).
- All measurements need to be possible to be conducted in parallel (non-adaptivity).



Pooled data



- Special case of compressed sensing (e.g. Donoho)
- In this talk: $d = 1$, *Quantitative Group Testing*
- QGT studied since the 1960's (Erdős, Rényi, Soderberg, Shapiro, Djackov, Kucherov, Gebrinski, ...) and of interest today (Alaoui et al., Feige & Lellouche, Gebhard et al., Karimi et al., Scarlett & Cevher, ...)

Sparsity

Definition

$\boldsymbol{\sigma}$ is sparse if
$$||\boldsymbol{\sigma}||_0 := k \ll n$$

- We assume $k = n^\theta$ for some $\theta \in (0, 1)$.
- Important in inference problems: e.g. compressed sensing is efficiently solvable by convex optimisation if the signal is sparse ($\ell_0 - \ell_1$ equivalence, Donoho 2013).

Theorem

If $\boldsymbol{\sigma}$ is sparse and A a Rademacher matrix or a Gaussian matrix, we can reconstruct $\boldsymbol{\sigma}$ from $A\boldsymbol{\sigma}$ efficiently by solving
$$\min ||z||_1 \qquad s.t. \qquad A\boldsymbol{\sigma} = z, z \in \mathbb{R}.$$

Near optimal

Lemma (Folklore lower bound)

The number of measurements m required for recovery of $\boldsymbol{\sigma}$ is at least

$$m \geq k \frac{\log(n/k)}{\log k} = \frac{\theta}{1-\theta} k.$$

- The number of possible results is $(k+1)^m$ and we need to distinguish $\binom{n}{k}$ possible ground-truth values.

Near optimal

Lemma (Djackov's lower bound)

The number of measurements m required for recovery of $\boldsymbol{\sigma}$ is at least

$$m \geq 2\frac{\theta}{1-\theta}k.$$

Near optimal

Lemma (Exponential time upper bound)

There is a simple randomised construction on

$$m \approx 2\frac{\theta}{1-\theta}k$$

measurements that allows exhaustive search to reconstruct $\boldsymbol{\sigma}$ w.h.p..

- Independent proofs by Feige & Lellouche and Gebhard et al.
- Simple: Any measurement chooses $n/2$ items uniformly at random.

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements)
- Irregular sparse parity check codes (Karimi et al.)
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.)
- Thresholding algorithms (Gebhard et al.)
- SubsetSelect Problem (Feige & Lellouche)

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements): requires $m = \Omega(k \log(n))$.
- Irregular sparse parity check codes (Karimi et al.)
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.)
- Thresholding algorithms (Gebhard et al.)
- SubsetSelect Problem (Feige & Lellouche)

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements): requires $m = \Omega\left(k \log(n)\right)$.
- Irregular sparse parity check codes (Karimi et al.): requires $m = \Omega\left(k \log(n)\right)$.
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.)

- Thresholding algorithms (Gebhard et al.)
- SubsetSelect Problem (Feige & Lellouche)

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements): requires $m = \Omega(k \log(n))$.
- Irregular sparse parity check codes (Karimi et al.): requires $m = \Omega(k \log(n))$.
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.): requires $m = \Omega(k \log(n))$.
- Thresholding algorithms (Gebhard et al.)
- SubsetSelect Problem (Feige & Lellouche)

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements): requires $m = \Omega(k \log(n))$.
- Irregular sparse parity check codes (Karimi et al.): requires $m = \Omega(k \log(n))$.
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.): requires $m = \Omega(k \log(n))$.
- Thresholding algorithms (Gebhard et al.): requires $m = \Omega(k \log(n))$.
- SubsetSelect Problem (Feige & Lellouche)

How to do it efficiently?

- Compressed Sensing (Basis Pursuit and refinements): requires $m = \Omega\left(k \log(n)\right)$.
- Irregular sparse parity check codes (Karimi et al.): requires $m = \Omega\left(k \log(n)\right)$.
- Binary group testing (e.g. Aldridge et al., Coja-Oghlan et al.): requires $m = \Omega\left(k \log(n)\right)$.
- Thresholding algorithms (Gebhard et al.): requires $m = \Omega\left(k \log(n)\right)$.
- SubsetSelect Problem (Feige & Lellouche): requires $m = \Omega\left(k \log(n)\right)$.

How to do it nearly optimal and efficiently?

Theorem (HKN2021+)

There is a randomised polynomially time construction coming with a polynomial-time inference algorithm that allows reconstruction of $\boldsymbol{\sigma}$ by no more than

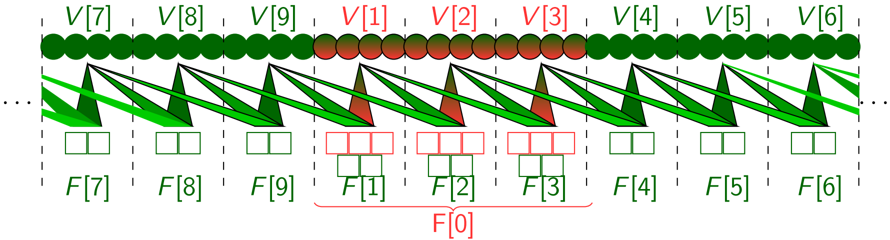
$$m = (4+\delta)\frac{1+\sqrt{\theta}}{1-\sqrt{\theta}}\left(2\frac{1-\theta}{\theta}k\right) = O(k)$$

measurements.

- Closing the previously conjectured $\log n$ gap up to a moderate multiplicative constant.
- Basic idea: Equip a clever version of Gebhard et al.'s thresholding algorithm with a spatially coupled pooling design.

Spatial Coupling

- Was invented in coding theory (Kukedar et al. 2013)
- Asymptotically vanishing *seed*
- Most of items are in the so-called *bulk*

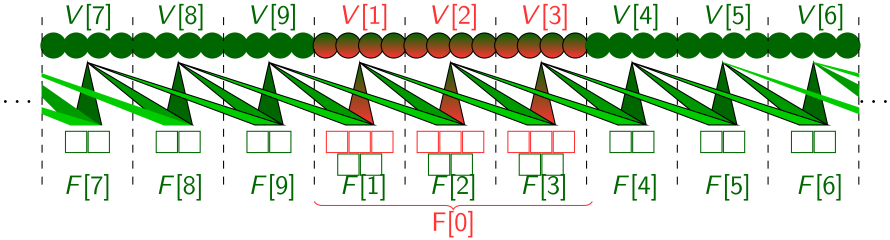


Decoding the seed

- The seed contains roughly $n' \approx \frac{n}{\sqrt{k}}$ items out of which $k' \approx \sqrt{k}$ have weight one.
- Apply an algorithm of your choice requiring $\approx k' \log(n') = o(k)$ measurements (we used Basis Pursuit).

Decoding the bulk

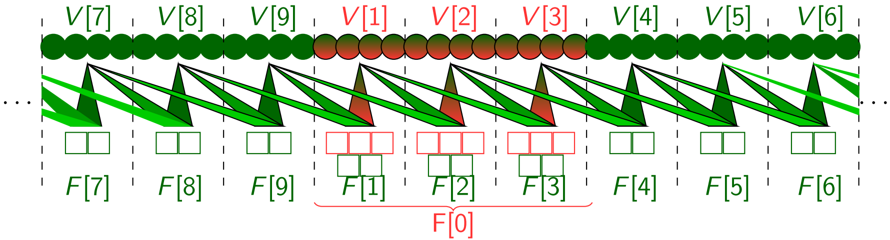
- Suppose we already decoded compartments $1 \dots i-1$ correctly.
- The *unexplained neighbourhood sum* of an item is the sum over its measurements subtracted by the weights of already contained items.
- The unexplained neighbourhood sum (random, binomially distributed) is increased by $\deg(x)$ if the weight of x is one.



Decoding the bulk

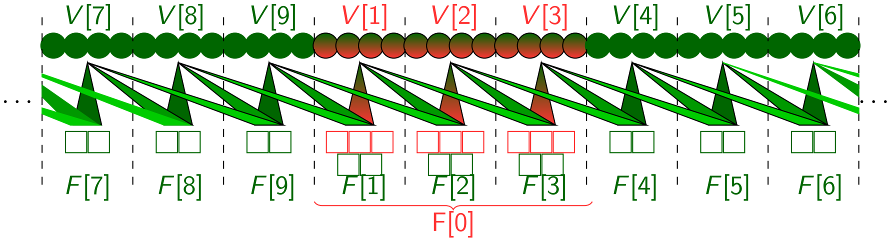
- Information in *close* compartments is much more valuable (weigh close compartments more in the sum).
- The summands of close compartments are significantly smaller. $\Rightarrow$ We need to normalise each summand!
- Instead of calculating $\boldsymbol{U}_x = \sum_{r=1}^{s} \boldsymbol{U}_x^j$ (the unexplained neighbourhood sum) we calculate

$$\boldsymbol{N}_x = \sum_{r=1}^{s} \omega_r \frac{\boldsymbol{U}_x^j - \mathbb{E}\left[\boldsymbol{U}_x^j\right]}{\sqrt{\mathrm{Var}\left(\boldsymbol{U}_x^j\right)}}.$$



Decoding the bulk

- This weighted unexplained normalised neighbourhood sum is still increased by a constant (depending on $\deg(x)$) if the weight of x is one.
- If enough measurements are conducted, the distributions between items of weight zero and weight one are well separated w.h.p..



Summary

- Spatial coupling was previously used to optimise constants (Coja-Oghlan et al., 2021).
- We used it to decrease the order of measurements.
- Simple thresholding is not enough (this only improved the constant) - normalised quantities allowed us to reduce the order.
- We could not use the (information-theoretically optimal) design with measurements of size $n/2$ as error terms in concentration results became too high.
- In the used design, any algorithm would require

$$m \geq 8 \frac{1-\theta}{\theta} k$$

measurements.

Thank you!

Questions?

... and (hopefully) answers!