Near optimal efficient decoding from pooled data

Max Hahn-Klimroth MAXIMILIAN.HAHNKLIMROTH@TU-DORTMUND.DE
TU Dortmund University, Faculty of Computer Science, Germany

Noela Müller N.S.MULLER@TUE.NL
Eindhoven University of Technology, Department of Mathematics and Computer Science, the Netherlands

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Consider n items, each of which is characterised by one of $d+1$ possible features in $\{0,\dots,d\}$. We study the inference task of learning these types by queries on subsets, or pools, of the items that only reveal a form of coarsened information on the features - in our case, the sum of all the features in the pool. This is a realistic scenario in situations where one has memory or technical constraints in the data collection process, or where the data is subject to anonymisation. Related prominent problems are the quantitative group testing problem, of which it is a generalisation, as well as the compressed sensing problem, of which it is a special case.

In the present article, we are interested in the minimum number of queries needed to efficiently infer the features, in the setting where the feature vector is chosen uniformly while fixing the frequencies, and one of the features, say 0, is dominant in the sense that the number $k = n^\theta, \theta \in (0,1)$, of non-zero features among the items is much smaller than n. It is known that in this case, all features can be recovered in exponential time using no more than $O(k)$ queries. However, so far, all *efficient* inference algorithms required at least $\Omega(k \ln n)$ queries, and it was unknown whether this gap is artificial or of a fundamental nature. Here we show that indeed, the previous gap between the information-theoretic and computational bounds is not inherent to the problem by providing an efficient algorithm that succeeds with high probability and employs no more than $O(k)$ measurements. This also solves a prominent open question for the quantitative group testing problem.

Keywords: Statistical inference, Quantitative group testing, Compressed Sensing, Pooled Data, Spatial Coupling

1. Introduction

Imagine a population of n items, each of which is characterised by one of finitely many distinct features, such as a class label in an object detection task, an age group, gender, or blood type. We are interested in inferring these types, using only a small number of coarse measurements on subgroups of the items. This problem, as introduced by Wang et al. (2016), is known as the *pooled data problem*. While its general framework is of relevance in many practical situations, a particularly prominent and topical instance of the pooled data problem is given by the quantitative group testing problem (Djackov, 1975; Gebhard et al., 2022; Grebinski and Kucherov, 2000; Karimi et al., 2019). In this case, the population is split into two types, which we interpret as *healthy* and *defective*, and the goal is to identify which individuals are the defective ones, by using tests that only provide the total number of defectives in the pooled subgroup. Other applications of the pooled data problem include DNA screening (Sham et al., 2002), traffic monitoring (Wang et al., 2015), machine learning (Liang and Zou, 2021; Martins et al., 2014) and signal recovery (Mazumdar and Pal, 2022).

In the following, we will assume that the labels of the items are chosen uniformly given their frequencies. Since the information on the features increases with the number of queries asked on them, a natural and important question concerns the minimum number of queries that are needed to successfully identify all the labels with high probability[1] over the choice of the labels. To address it, one can derive both upper and lower bounds: A sequence m_0 such that for $m \geq m_0$, the probability of making an error does not tend to one is called an *information-theoretic lower bound*. On the other hand, a sequence m_0 such that for $m \geq m_0$, *exponential time* algorithms like exhaustive search are guaranteed to recover all features w.h.p. is called an *information-theoretic upper bound*.

Having obtained information-theoretic bounds and thus identified a regime where inference is theoretically possible, a second important question is concerned with the minimal m for which all features can be recovered *efficiently* (in polynomial time).

Our article addresses this second question in the case where one of the features is dominant. For this setting, we provide an efficient algorithm that w.h.p. infers all labels correctly, while using no more than $O(k)$ queries. This algorithm is the first one to match the information-theoretically optimal order of queries. In the special case of quantitative group testing, this result resolves the basic, yet open question whether efficient, information-theoretically optimal inference is possible. We continue to describe the precise model and related results.

1.1. Model and terminology

Consider n items $x_1, \dots, x_n$, each of which is assigned a label $\boldsymbol{\sigma}_i := \boldsymbol{\sigma}(x_i) \in \{0, 1, 2, \dots, d\}$. The vector $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n)$ of item-labels constitutes the *ground-truth* or *signal* that we aim to infer by performing m queries on (multi-)subsets $a_1, \dots, a_m$ of the items, which we call *pools*. There is freedom both in the choice of the pools $a_1, \dots, a_m$ as well as in the nature of the queries. In the design of the pools, we restrict ourselves to the *non-adaptive* setting, where all m pools have to be constructed before conducting any queries. This constraint is predominant in theoretical work on the quantitative group testing problem and of relevance in practical applications due to scalability and stability considerations Zhou et al. (2014). With respect to the queries, we consider the *additive model*, where for each $i = 1, \dots, m$, the total *weight* $\hat{\boldsymbol{\sigma}}_i := \sum_{j : x_j \in a_i} \boldsymbol{\sigma}_j$ of pool i is measured. In particular, this model is a natural extension of the quantitative group testing problem, where the measurement $\hat{\boldsymbol{\sigma}}_i$ corresponds to the number of defectives in the $i-$th pool. On the other hand, it reveals less information on the labels than other commonly studied variants of the problem, where one measures a histogram of the frequencies of each label within the given pool (see El Alaoui et al. (2019); Scarlett and Cevher (2017)).

It is well-known that the presence of a dominant label, say 0, is a distinguishing feature in the theoretical analysis of the model. Denote by $k_0, \dots, k_d$ the numbers of items with label $0, \dots, d$, respectively, and by $k := \sum_{i=1}^{d} k_i$ the total number of non-zero labels. If $k/n \to \alpha \in (0, 1)$ as $n \to \infty$, the problem is called *linear pooled data problem*, while the *sublinear pooled data problem* considers $k \sim n^{\theta}$ for $\theta \in (0, 1)$. In this article, we study the sublinear regime. One reason behind the interest in this regime is its practicability. In the context of epidemiology, early numbers of defectives can be captured by this setting according to Heaps' Law (Wang et al., 2011). In the context of machine learning, pooled measurements have been applied to image moderation tasks (Liang and Zou, 2021), where the sublinear regime corresponds to the detection of rare, but inappropriate images.

1. With high probability (w.h.p.) means with probability tending to 1 as $n \to \infty$.

Within the sublinear setting, we assume that $k_0, \dots, k_d$ are known a priori, for example through empirical findings, and that for each label $i = 1, \dots, d$, there exists $\varepsilon_i = \Theta(1)$ such that $k_i = \varepsilon_i k$. This model choice is analogous to parametrisations in the linear regime as in El Alaoui et al. (2019). Finally, we perform an average-case analysis of the problem and from here on, we will thus assume that the ground-truth $\boldsymbol{\sigma}$ is chosen uniformly at random among all vectors having exactly k_i entries of type i, where $i = 0, \dots, d$. To realise this assumption in practice, one can apply a uniform permutation to the items before running any inference algorithm.

In the next two subsections, we briefly review previous work on the problem and then give an overview of our main results. After fixing the notation, Section 2 describes our pooling scheme, while Section 3 is devoted to the presentation of the inference algorithm. The final Section 4 contains a summary of our findings.

1.2. Information-theoretic and computational bounds

Information-theoretic lower bounds In the special case of the quantitative group testing problem, Djackov (1975) provides the explicit information-theoretic lower bound $m_{\mathrm{QGT}} \geq 2\frac{1-\theta}{\theta}k$.

Turning to the general case, a meaningful information-theoretic lower bound m_{count} can be obtained with the help of a simple counting argument: For fixed $k_0, \dots, k_d$, there are $\binom{n}{k_0, k_1, \dots, k_d}$ different values that the ground-truth $\boldsymbol{\sigma}$ can take. On the other hand, each query outputs a value between 0 and $W := \sum_{j=1}^{d} jk_j$. For unambiguous inference to be possible with high probability, we thus need $\liminf_{n\to\infty} (W+1)^{m_{\mathrm{count}}} / \binom{n}{k_0, k_1, \dots, k_d} \geq 1$. In our setting, where $k = n^\theta$ for $\theta \in (0,1)$, this argument has the consequence that for fewer than $m_{\mathrm{count}} = \Omega(k)$ pools, there is no hope to successfully reconstruct $\boldsymbol{\sigma}$ w.h.p.

Information-theoretic upper bounds By means of a non-constructive argument, Grebinski & Kucherov (Grebinski and Kucherov, 2000) show that

$$m_{\mathrm{GK}} = 4W \ln\left(\frac{n}{W} + 1\right) \ln^{-1}(W) = O(k)$$

pools suffice to reconstruct $\boldsymbol{\sigma}$ w.h.p. Moreover, in the special case of the quantitative group testing problem, the leading constant was further reduced by a factor of 2 in independent works of Gebhard et al. (2022) and Feige and Lellouche (2020). Since these results asymptotically match the information-theoretic lower bound, the pooled data problem can be regarded as almost understood from an information-theoretic point of view. However, with respect to efficient algorithms, the picture is a different one.

Efficient Algorithms Natural candidates for efficient inference algorithms in the pooled data problem are those that also apply to the more general sparse compressed sensing problem. The literature on this topic is vast, with the foundational contributions of Candes et al. (2006) and Donoho (2006), but we refrain from reviewing it in more detail at this point. Linear programming techniques from this context guarantee successful inference w.h.p. with $\Theta(k \ln n)$ measurements in the sublinear regime.

In the extensively studied quantitative group testing problem, one might hope that taking into account the specific structure of the problem yields some improvement on the order of the pools needed. And indeed, more specialised algorithms exist, see e.g. Coja-Oghlan et al. (2020); Feige and Lellouche (2020); Gebhard et al. (2022); Karimi et al. (2019). However, even these ideas

have failed to beat the lower bound of $\Omega(k \ln n)$. Thus, when comparing what is information-theoretically achievable and what is efficiently achievable, we are faced with a multiplicative gap of order $\ln n$, for all values of d.

1.3. Main result

In this article, we propose and analyse an efficient algorithm that reconstructs the ground truth $\boldsymbol{\sigma}$ correctly w.h.p., while using no more than $O(k)$ measurements, and thereby overcome the multiplicative $\ln n$ gap between the current algorithmic and information-theoretic bounds. Our algorithm makes use of a random pooling design that is based on the so-called *spatial coupling*-technique from coding theory (Felström and Zigangirov, 1999; Kudekar et al., 2011, 2013). In particular, all pools have the same fixed, non-random size. The inference algorithm given the pooling scheme is then based on a thresholding idea.

Theorem 1 *Let $d \in \mathbb{N}$, $\theta \in (0,1)$ and abbreviate $k = k(n) = n^\theta$. Fix a sequence of proportions $(\varepsilon_1, \dots, \varepsilon_d) = (\varepsilon_1(n), \dots, \varepsilon_d(n)) \in (0,1)^d$ with $(\varepsilon_1, \dots, \varepsilon_d) = \Theta(1)$ and $\sum_{w=1}^d \varepsilon_w = 1$, and choose $\boldsymbol{\sigma} \in \{0,1,\dots,d\}^n$ uniformly among all vectors having exactly $\varepsilon_w k$ entries of value w for $w = 1, \dots, d$. Then for each $\delta > 0$, there are a randomised, non-adaptive poly(n)-time construction of a pooling scheme and a deterministic poly(n)-time algorithm that jointly allow to recover $\boldsymbol{\sigma}$ w.h.p. within the additive model, while using no more than*

$$m_{PD} = (8+\delta)\,\frac{1+\sqrt{\theta}}{1-\sqrt{\theta}}\left(\sum_{w=1}^{d} w^2 \varepsilon_w\right)\frac{1-\theta}{\theta}k$$

pools.

Most notably, when applied to the *quantitative group testing problem*, Theorem 1 guarantees the existence of a polynomial-time construction of a testing scheme coming with a polynomial-time algorithm that recovers $\boldsymbol{\sigma}$ w.h.p. using no more than

$$m_{\mathrm{SC}} = (8+\delta)\frac{1+\sqrt{\theta}}{1-\sqrt{\theta}}\,\frac{1-\theta}{\theta}k$$

tests. Thus, the performance of our algorithm matches the information-theoretic lower bound m_{QGT} up to a moderate constant.

1.4. Related problems

Problems related to the pooled data problem arise in a variety of contexts and have been of fundamental interest to mathematicians for a long time. For example, extensions of the quantitative group testing problem ($d = 1$) have been studied since the 1960's by, among others, Erdős and Rényi (1963), Djackov (1975) and Shapiro (1960). El Alaoui et al. (2019) and Wang et al. (2016) introduce a variant where $\boldsymbol{\sigma}$ is is a vector in $\{0,1,\dots,d\}^n$ and queries output the numbers of items of each label within the pool, while Bshouty (2009) studies the *coin weighing problem*, where every query returns the sum of the contained labels. Clearly, any algorithm that uses the coarser information from the setting of Bshouty (2009) can also be used for inference within the framework of El Alaoui et al. (2019); Wang et al. (2016). Finally, the pooled data problem can be seen as a special

case of the *compressed sensing problem* (Candes et al., 2006; Donoho and Tanner, 2006; Foucart and Rauhut, 2013), which generally asks for recovery of a high-dimensional signal $\boldsymbol{\sigma} \in \mathbb{R}^n$ from a small number of linear measurements on its components.

A pooling scheme of a similar design as the one underlying Theorem 1 in combination with a thresholding algorithm has been used in recent work on the binary group testing problem (Coja-Oghlan et al., 2020). In this problem, which differs from the quantitative group testing problem, tests do not output the number of defectives within a given pool, but simply the binary information whether a given pool contains a defective individual. We will discuss the similarities to and differences from the approach in Coja-Oghlan et al. (2020) in the discussion after our algorithm has been presented in detail.

2. Model

2.1. Getting started

Generally, we write $[a]$ for the set $\{1, \dots, a\}$ and if $z \in \mathbb{R}^p$, we also use the notation $z(t)$ to refer to the t-th component of z, where $t \in [p]$.

Recall that we aim to infer the labels $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n \in \{0, 1, \dots, d\}$ of n items $x_1, \dots, x_n$ by measuring label-sums in m multi-subsets $a_1, \dots, a_m$ of $\{x_1, \dots, x_n\}$. We call these multi-subsets pools and assume that the vector $\boldsymbol{\sigma} := (\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n) \in \{0, 1, \dots, d\}^n$ is chosen uniformly at random from all vectors containing exactly k_i entries of value i for each $i \in [d]$. Here, $k_i = \varepsilon_i n^\theta$ for $\varepsilon_i = \Theta(1)$, $i \in [d]$, and $\theta \in (0, 1)$. We abbreviate $k = \sum_{i=1}^d k_i = n^\theta$. Finally, $W = \sum_{i=1}^d i k_i$ denotes the total weight of $\boldsymbol{\sigma}$.

In the following, we will represent pooling schemes $\{a_1, \dots, a_m\}$ as bipartite multi-graphs. In this representation, the pools and the items yield the two vertex classes of the bipartite graph (see Figure 1). An edge in the graph is present whenever the incident item is an element of the incident pool. This prescription yields a multi-graph since in our scheme, items will be allowed to appear multiple times in a given pool. We denote the neighbourhood of item x_i in this graph by ∂x_i and the neighbourhood of pool a_j by ∂a_j. These are understood to be multi-sets of pools and items, respectively. If we work with sets rather than multi-sets, we use the notation $\partial^\star x_i, \partial^\star a_j$ for the sets obtained from ∂x_i and ∂a_j. Finally, we denote the label-sum of pool a_j by $\hat{\boldsymbol{\sigma}}_j$ such that $\hat{\boldsymbol{\sigma}}_j = \sum_{i: x_i \in \partial a_j} \boldsymbol{\sigma}_i$.

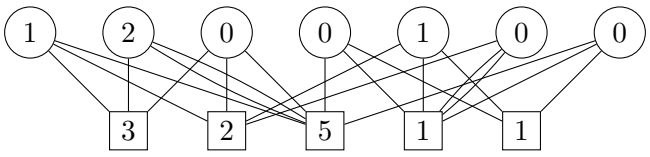

Figure 1: Graphical representation of a pooling scheme: Here, the $n = 7$ items are represented by circles, while the $m = 5$ pools are represented by squares. Edges between items and pools are present whenever an item is an element of the corresponding pool.

2.2. The pooling scheme

The polynomial-time pooling scheme that constitutes the basis for our algorithm builds upon the introduction of a "spatial" order to both items and pools. To this end, we introduce a parameter

$\varepsilon > 0$ that will later be chosen small enough and depending on δ from Theorem 1. We then partition $V := \{x_1, \dots, x_n\}$ into ℓ *compartments* $V[s], \dots, V[s+\ell-1] \subset V$ of (almost) equal sizes $|V[i]| \in \{\lfloor n/\ell \rfloor, \lceil n/\ell \rceil\}$, where $\ell = \lceil k^{1/2-\varepsilon} \rceil$ and $s = \lceil \ell^{1/2-\varepsilon/2} \rceil$. This partition will allow us to successively infer the labels of each compartment, where we proceed from s to $s+\ell-1$ and use the information of previous compartments along the way. However, to get this idea started properly, the first few compartments need some extra attention.

We facilitate the initial steps of the algorithm by the introduction of $s-1$ artificial compartments $V[1], \dots, V[s-1]$ (this also explains why the labelling in the previous paragraph starts at s). These contain $n' = (s-1)\lceil n/\ell \rceil$ *auxiliary items* that are distributed equally among the $s-1$ compartments. To equip these auxiliary items with labels that behave as the original variable labels, we sample an assignment $\boldsymbol{\tau} \in \{0, 1, \dots, d\}^{n'}$ uniformly at random from all vectors with exactly $k'_i = \lceil (s-1)\varepsilon_i k \ell^{-1} \rceil$ items of weight $i \in [d]$. This only takes polynomial time and in particular, $\boldsymbol{\tau}$ is known. In the remainder of this article, we call the $s-1$ compartments $V_{\text{seed}} = V[1] \cup \dots \cup V[s-1]$ the *seed*, while the remaining compartments ℓ constitute the *bulk* V_{bulk}.

Analogously to the partition of the bulk, for an integer m divisible by $\ell + s - 1$, we divide the m pools into $\ell + s - 1$ many compartments $F[1], \dots, F[\ell+s-1]$ such that each of the compartments contains exactly $m/(\ell+s-1) \sim m/\ell$ pools.

After this partitioning, items for the pools are chosen as follows: First, let

$$\Gamma := \frac{ns}{\sqrt{m}(\ell+s-1)} + O(s)$$

be an integer divisible by s, which will be the number of items in each pool. Let $j \in [\ell+s-1]$ be the index of one of the pool compartments. Then each pool $a \in F[j]$ chooses its Γ items exclusively from the s "previous" compartments $V[j-(s-1)], \dots, V[j]$, where it selects exactly Γ/s items from each of these compartments uniformly at random with replacement. Here and in the following, for $r = 0, \dots, s-2$, we identify $V[-r]$ with $V[\ell+s-1-r]$ as well as $F[\ell+s+r]$ with $F[r+1]$, which equips the random bipartite graph with a ring structure. In particular, the number s is called the *sliding window*. The terminology and construction are illustrated by Figure 2.

As described above, this pooling scheme gives rise to a bipartite multi-graph. We denote the (random) degrees of the items $x_1, \dots, x_n$ in this graph by $\boldsymbol{\Delta}_{x_1}, \dots, \boldsymbol{\Delta}_{x_n}$. The number of neighbours of item x_i among the elements of $F[i+j]$ is denoted by $\boldsymbol{\Delta}_{x_i}[j]$, where $j = 0, \dots, s-1$. Furthermore, as the pools sample their items with replacement, we introduce $\boldsymbol{\Delta}^\star_{x_i}, \boldsymbol{\Delta}^\star_{x_i}[j]$ as the corresponding numbers of *distinct* pools that item x_i participates in. As we will see later, the effect of multi-edges is almost negligible, as $\boldsymbol{\Delta}^\star_{x_i} = (1-o(1))\boldsymbol{\Delta}_{x_i}$ w.h.p., but their presence simplifies the analysis mildly. Finally, set $\Delta := \mathbb{E}[\boldsymbol{\Delta}_{x_i}]$ and $\Delta^\star := \mathbb{E}\left[\boldsymbol{\Delta}^\star_{x_i}\right]$.

3. Algorithm outline

Motivation As a starting point for the algorithm, we consider the influence of an arbitrary variable $x \in V$ on the queries on the pools that it participates in.

While a change in the label of x typically only marginally affects the outcome of the query at one of its neighbouring pools, the situation is different if we consider all $\boldsymbol{\Delta}_x$ neighbours of x jointly. Indeed, this number is binomially distributed and thus concentrated around its mean $\Delta \sim \sqrt{m}s/\ell$. More precisely, the Chernoff bound guarantees that, w.h.p.,

$$\boldsymbol{\Delta}_x = \Delta \pm \ln n \sqrt{\Delta} = (1+o(1))\Delta.$$

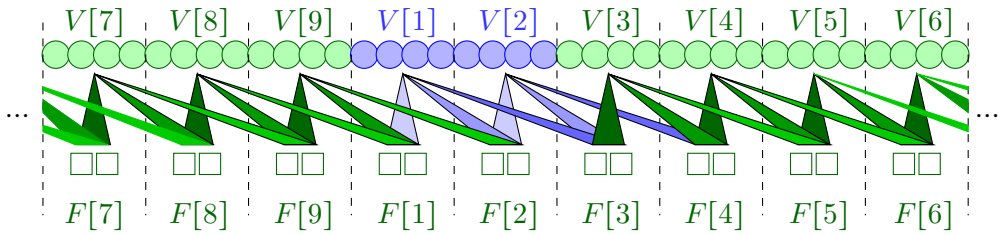

Figure 2: Schematic representation of the pooling scheme with $n = 28$ items, $\ell = 7$ compartments, 18 measurements and a sliding window of size $s = 3$. The number of (blue) auxiliary items, whose weight is known a priory, is 8 in this example. The figure is an adaptation of Figure 2 in Coja-Oghlan et al. (2020).

By definition of s, ℓ and for $m = \Omega(k)$ (in agreement with the information-theoretic lower bound), we have $\Delta = \Omega(k^{1/4+\varepsilon^2/2})$. Therefore, if we define the *neighbourhood sum* of x as the sum of all the outcomes of pools in which x occurs, and alter the label of x, the change in the neighbourhood sum is of order $\Omega(k^{1/4+\varepsilon^2/2})$. As a consequence, the neighbourhood sum of x is highly dependent on $\boldsymbol{\sigma}_x$.

In light of this observation, it is natural to try to discern the labels by thresholding neighborhood sums, provided that these are concentrated well enough for items of different types and that sufficiently many measurements are conducted. This idea, which does *not* make use of our sophisticated pooling scheme, has previously been applied to the quantitative group testing problem (Gebhard et al., 2022), where the authors show that it leads to successful recovery of $\boldsymbol{\sigma}$ using $\Theta(k \ln n)$ measurements w.h.p.

We aim to improve this result through the spatially coupled design, which provides additional information on the labels at each inference step. For simplicity, we assume from now on that $x \in V[s]$ is an item from the first bulk compartment. Then the pooling scheme in combination with the auxiliary compartments ensures that:

- Item x is *only* contained in pools from the s compartments $F[s], \dots, F[2s-1]$.
- For each $j = 0, \dots, s-1$, any pool from compartment $F[s+j]$ contains a proportion of $1-(j+1)/s$ of already known labels (namely, those from compartments $V[1], \dots, V[s-1]$).

Therefore, rather than to simply sum up the labels of items that share a pool with x, one should construct the neighbourhood sum and then subtract all contributions from known or, more generally, previously inferred labels in order to lay bare the influence of the unknown labels. We call the reduced sum arising from this idea the *unexplained neighbourhood sum* of x.

Unexplained neighbourhood sums already yield some improvement over naive neighbourhood sums. However, the second observation above also indicates that pools from compartments close to $F[s]$ actually reveal *more* information on the items in $V[s]$ than pools from far apart compartments, since they contain a larger portion of known labels. For example, the pools in $F[s]$ have unexplained neighbourhood sums that exclusively involve labels from $V[s]$. One idea to incorporate this imbalance between the information coming from the different compartments and to further improve the concept of an unexplained neighbourhood sum is to introduce weights $\omega_1, ..., \omega_s \in (0, 1]$ to scale

the contributions accordingly. We call such a compartment-wise linear combination of unexplained neighbourhood sums *weighted unexplained neighbourhood sum*[2].

Alas, it turns out that again, this process does not yield the desired improvement. This is due to the fact that while the information from close compartments is indeed much more valuable, on the other hand the expected size of the unexplained neighbourhood sum is by a factor of $s = n^{\Omega(1)}$ larger in the farthest away compartment than in the closest one. Thus, despite the contrary effort, the influence of the first compartment almost vanishes in the weighted unexplained neighbourhood sum. To compensate for this effect, we normalise the unexplained neighbourhood sum in each compartment and sum up those normalised quantities in the described weighted fashion with the *weighted normalised unexplained neighbourhood sum* $\mathcal{N}_x$, which is the core quantity of our algorithm.

Key quantities In this subsection, we formalise the notions of the last subsection. Let $\tilde{\sigma} \in \{0,\dots,d\}^n$ be the current estimate of $\boldsymbol{\sigma}$ during any step of the algorithm. Then for each item $x \in V[i]$, $i = s,\dots,\ell+s-1$, and $j = 0,\dots,s-1$ we first define the *unexplained neighbourhood sum* of x into compartment $F[i+j]$ with respect to the estimate $\tilde{\sigma}$ as

$$\mathcal{U}_x^j := \mathcal{U}_x^j(\hat{\boldsymbol{\sigma}},\tilde{\sigma}) = \sum_{a\in\partial^\star x\cap F[i+j]}\left(\hat{\boldsymbol{\sigma}}_a - \sum_{p=1}^{s-j-1}\sum_{y\in\partial a\cap F[i-p]}\tilde{\sigma}_y\right). \tag{3.1}$$

We illustrate this random variable for the special case $d = 1$. In this case, given the pool design, and under the assumption of perfect knowledge of $\boldsymbol{\sigma}$,

$$\mathcal{U}_x^j(\hat{\boldsymbol{\sigma}},\boldsymbol{\sigma}) \approx \mathrm{Bin}\left((j+1)\boldsymbol{\Delta}_x^\star[j]\frac{\Gamma}{s},\frac{k}{n}\right) + \frac{\boldsymbol{\Delta}_x}{s}\boldsymbol{\sigma}_x.$$

Indeed, x has $\boldsymbol{\Delta}_x^\star[j]$ neighbours in compartment $F[i+j]$, each of which features roughly $(j+1)\Gamma/s$ so far unexplained items. This is only a heuristic approximation of the unexplained neighbourhood sum, but still instructive. The precise distributions of the unexplained neighbourhood sums throughout the inference process are derived in Lemma 3.

In a next step, we define the *normalised unexplained neighbourhood sum* $\mathcal{N}_x^j$ of x into compartment $F[i+j]$. As explained in the previous paragraph, this quantity takes (an approximation of) the expectation and the variance of $\mathcal{U}_x^j$ into account. Under the assumption that the current estimate of the ground truth has been mostly correct on the previous compartments, we define an estimate of $\mathbb{E}[\mathcal{U}_x^j|\boldsymbol{\Delta}_x[j],\boldsymbol{\Delta}_x[j]^\star]$ by setting

$$M_x^j := \sum_{w=1}^{d} w\frac{k_w}{n}\left((j+1)\boldsymbol{\Delta}_x^\star[j]\frac{\Gamma}{s} - \boldsymbol{\Delta}_x[j]\right).$$

The conditional variance of $\mathcal{U}_x^j$ is, with high probability, $(1+o(1))(j+1)k^{2\varepsilon}$ by the choice of Γ and the concentration properties of $\boldsymbol{\Delta}_x[j],\boldsymbol{\Delta}_x^\star[j]$. The normalised unexplained neighbourhood sum $\mathcal{N}_x^j$ of x into compartment $F[i+j]$ is then defined as

$$\mathcal{N}_x^j := \frac{\mathcal{U}_x^j - M_x^j}{\sqrt{(j+1)k^{2\varepsilon}}}.$$

2. A weighted unexplained neighbourhood sum is the key quantity in the analysis of optimal binary group testing (Coja-Oghlan et al., 2020).

Through $\mathcal{U}_x^j$, the sum $\mathcal{N}_x^j$ depends on $\boldsymbol{\sigma}_x$. Indeed, it is approximately of order

$$\mathcal{N}_x^j(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}) \approx \boldsymbol{\sigma}_x \Delta s^{-1} \sqrt{(j+1)k^{2\varepsilon}}^{-1} = C\boldsymbol{\sigma}_x \sqrt{j+1}^{-1}$$

for some constant $C > 0$. But, in contrast to $\mathcal{U}_x^j$, it is comparable between items in *close* compartments and *far apart* compartments.

Finally, the *weighted normalised unexplained neighbourhood sum* of item x with a specified choice of weights is defined as

$$\mathcal{N}_x := \sum_{j=0}^{s-1} (j+1)^{-0.5} \mathcal{N}_x^j. \tag{3.2}$$

Finding thresholds The idea of the inference algorithm is quite simple: We define thresholds $T^{0,1}, T^{1,2}, \dots, T^{d-1,d}$ such that item x is classified as having weight $i \in [d-1]$ if $T^{i-1,i} < \mathcal{N}_x \leq T^{i,i+1}$, as having label 0 if $\mathcal{N}_x \leq T^{0,1}$ and as having label d otherwise. With respect to the information-theoretic lower bound m_{QGT} in the quantitative group testing problem, we assume that the spatially coupled pooling scheme involves

$$m = 2c\frac{1-\theta}{\theta}k$$

pools for some constant $c = c_{d,\theta} \geq 1$ that may depend on d and θ. We then define the thresholds $T^{i,i+1}$ for $i \in [d-1]$ and additionally $T^{0,1}$ as

$$T^{i,i+1} := \left(i + \frac{1}{2}\right)\sqrt{\frac{2c(1-\theta)}{\theta}} \ln s \qquad \text{and} \qquad T^{0,1} := \frac{1}{1+\sqrt{\theta}}\sqrt{\frac{2c(1-\theta)}{\theta}} \ln s.$$

The threshold $T^{0,1}$ takes a slightly different form, as it tells apart the much more numerous items of weight 0 from the items of non-zero weight. The choice of $T^{0,1}, \dots, T^{d-1,d}$ is explained in more detail in the full version (Hahn-Klimroth and Müller, 2021).

3.1. Algorithm

Recall $\boldsymbol{\tau} \in \{0, 1, \dots, d\}^{n'}$ from Section 2.2. Using the notation from the previous sections, our algorithm does the following.

Set $\tilde{\sigma} = (\boldsymbol{\tau}, 0) \in \{0, 1, \dots, d\}^{n'+n}$;
for $i = s, \dots, \ell + s - 1$ **do**
  For any individual $x \in V[i]$ calculate $\mathcal{N}_x$;
  Set $\tilde{\sigma}_x = \begin{cases} 0, & \text{if} \quad \mathcal{N}_x \leq T^{0,1} \\ d, & \text{if} \quad \mathcal{N}_x > T^{d-1,d} \\ b, & \text{if} \quad 1 \leq b \leq d-1 \text{ and } \mathcal{N}_x \in \left(T^{b-1,b}, T^{b,b+1}\right]. \end{cases}$
end

Algorithm 1: Algorithm for the pooled data problem with thresholds $T^{0,1}, \dots, T^{d-1,d}$.

It turns out that, for a not too large constant c, we find the following performance guarantee for Algorithm 1.

Proposition 2 *Let $\delta > 0$. Then $\varepsilon > 0$ can be chosen sufficiently small such that if*

$$c \geq \frac{(4+\delta)(1+\sqrt{\theta})}{1-\sqrt{\theta}} \sum_{w=1}^{d} w^2 \varepsilon_w,$$

the output $\tilde{\sigma}$ of Algorithm 1 coincides with $\boldsymbol{\sigma}$ w.h.p.

Outline of the proof of Proposition 2 We prove Proposition 2 in three steps. First, we find a reasonably well-behaved substitute $\boldsymbol{N}_x$ of $\mathcal{N}_x$ that uses idealised information and is therefore easier to study. Secondly, we bound the probabilities that thresholding $\boldsymbol{N}_x$ (had we access to it) would lead to a wrong classification of x. Finally, we prove Proposition 2 by an inductive argument that ensures that with sufficiently high probability, $\mathcal{N}_x$ is not too far from $\boldsymbol{N}_x$, for all x.

For step 1, we observe that the quantity $\mathcal{U}_x^j(\hat{\boldsymbol{\sigma}}, \tilde{\sigma})$ has a particularly accessible form, if the estimate $\tilde{\sigma}$ and the ground truth $\boldsymbol{\sigma}$ agree on all previously inferred compartments. We therefore introduce specific notation for the unexplained neighbourhood sum of x into compartment $F[i+j]$ with respect to the correct labels $\boldsymbol{\tau}, \boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_{n(i-1)/\ell}$. To this end, denote by $\boldsymbol{\sigma}^i$ the $(n'+n)$-dimensional vector which agrees with $(\boldsymbol{\tau}, \boldsymbol{\sigma})$ up to coordinate $n' + n(i-1)/\ell$ and with $\boldsymbol{\sigma}^i_b = 0$ for $b > n' + n(i-1)/\ell$. We then define

$$\boldsymbol{U}_x^j := \mathcal{U}_x^j(\hat{\boldsymbol{\sigma}}, \boldsymbol{\sigma}^i).$$

Thus, $\boldsymbol{U}_x^j$ is the (random) unexplained neighbourhood sum of item x into compartment $F[i+j]$ with respect to the true $\boldsymbol{\sigma}$. Moreover, denote by $\boldsymbol{k}_i^{(j)}$ the random number of items of weight i in compartment $j+s-1$ in the spatial coupling set-up and abbreviate

$$\underline{\boldsymbol{k}}^{(j)} := \left(\boldsymbol{k}_i^{(j)}\right)_{i=1\dots d} \qquad \text{and} \qquad \underline{\boldsymbol{k}} := \left(\underline{\boldsymbol{k}}^{(j)}\right)_{j\in[\ell]}.$$

Finally, let the σ-algebra $\mathcal{E}_x$ be

$$\mathcal{E}_x := \sigma\left(\partial x, \boldsymbol{\sigma}_x, \underline{\boldsymbol{k}}\right).$$

Based on the idealised information $\mathcal{E}_x$, we define the quantities

$$\boldsymbol{N}_x^j := \frac{\boldsymbol{U}_x^j - \mathbb{E}\left[\boldsymbol{U}_x^j \middle| \mathcal{E}_x\right] + \boldsymbol{\Delta}_x[j]\boldsymbol{\sigma}_x}{\sqrt{(j+1)k^{2\varepsilon}}} \qquad \text{as well as} \qquad \boldsymbol{N}_x := \sum_{j=0}^{s-1} (j+1)^{-0.5} \boldsymbol{N}_x^j, \tag{3.3}$$

Of course, if the estimate $\tilde{\sigma}$ has inferred every label correctly so far, then $\boldsymbol{U}_x^j$ agrees with $\mathcal{U}_x^j$. Unfortunately, the actual values of $\boldsymbol{\sigma}$ and $\boldsymbol{U}_x^j$ are unknown at any specific stage of the algorithm, and thus, we can neither compute $\boldsymbol{U}_x^j$ nor $\boldsymbol{N}_x$ exactly. However, the main strategy of the proof is to analyse $\boldsymbol{N}_x$ and then to show that w.h.p., the guess $\mathcal{N}_x$ is sufficiently close to $\boldsymbol{N}_x$ for the concentration properties of $\boldsymbol{N}_x$ to be transferred.

In Lemma 3, we analyse the distribution of each $\boldsymbol{U}_x^j$ given $\mathcal{E}_x$, which is basically a weighted sum over the coordinates of independent multinomially distributed random vectors. This distributional insight formalises the intuition that was given in the last section for the case $d = 1$. A detailed proof of Lemma 3 can be found in the full version of this article (Hahn-Klimroth and Müller, 2021).

Lemma 3 *Let $x \in V[i]$ and $j \in \{0, \dots, s-1\}$. Let*

$$\boldsymbol{X}_x^{(i:j)} := \mathrm{Mult}\left(\frac{\boldsymbol{\Delta}_x^\star[j]\Gamma}{s} - \boldsymbol{\Delta}_x[j], \frac{\boldsymbol{k}_1^{(j)} - \mathbf{1}\{\boldsymbol{\sigma}_x = 1\}}{n/\ell - 1}, \dots, \frac{\boldsymbol{k}_d^{(j)} - \mathbf{1}\{\boldsymbol{\sigma}_x = d\}}{n/\ell - 1}\right)$$

and

$$\boldsymbol{X}_x^{(r:j)} \sim \mathrm{Mult}\left(\frac{\boldsymbol{\Delta}_x^\star[j]\Gamma}{s}, \frac{\ell \boldsymbol{k}_1^{(r)}}{n}, \dots, \frac{\ell \boldsymbol{k}_d^{(r)}}{n}\right)$$

for $r = i+1, \dots, i+j$ be independent multinomial random variables given $\mathcal{E}_x$. Then

$$\boldsymbol{U}_x^j \stackrel{d}{=} \boldsymbol{\Delta}_x[j]\boldsymbol{\sigma}_x + \sum_{r=i}^{i+j}\sum_{w=1}^{d} w\boldsymbol{X}_x^{(r:j)}(w) \qquad \text{given } \mathcal{E}_x. \tag{3.4}$$

From Lemma 3, it is straightforward to calculate the conditional mean and variance of $\boldsymbol{U}_x^j$ given $\mathcal{E}_x$. In particular, it turns out that $\boldsymbol{N}_x$ given $\mathcal{E}_x$ can be written as a weighted sum over negatively associated Bernoulli random variables. Therefore, by a straightforward application of Bernstein's inequality, we find that the idealised weighted normalised unexplained neighbourhood sums $\boldsymbol{N}_x$ are tightly concentrated around their means.

Lemma 4 *For any $\alpha \in (0,1)$, set $T_\alpha^{0,1} := (1-\alpha)\sqrt{\frac{2c(1-\theta)}{\theta}}\ln s$. Then, there exists a constant $D > 0$ that depends on d and θ and a sequence $\zeta_n = o\left(n^{-2}\right)$ such that*

$$\mathbb{P}\left(\boldsymbol{N}_x > T_\alpha^{0,1}\middle|\boldsymbol{\sigma}_x = 0\right) \le Ds^{-\frac{(1-\alpha)^2 c(1-\theta)}{\theta\sum_{w=1}^d w^2\varepsilon_w}} + \zeta_n, \quad \mathbb{P}\left(\boldsymbol{N}_x \le T_\alpha^{0,1}\middle|\boldsymbol{\sigma}_x = 1\right) \le Ds^{-\frac{\alpha^2 c(1-\theta)}{\theta\sum_{w=1}^d w^2\varepsilon_w}} + \zeta_n$$

as well as

$$\mathbb{P}\left(\boldsymbol{N}_x > T^{i,i+1}\middle|\boldsymbol{\sigma}_x = i\right) \le Ds^{-\frac{c(1-\theta)}{4\theta\sum_{w=1}^d w^2\varepsilon_w}} + \zeta_n \quad \text{and}$$

$$\mathbb{P}\left(\boldsymbol{N}_x \le T^{i,i+1}\middle|\boldsymbol{\sigma}_x = i+1\right) \le Ds^{-\frac{c(1-\theta)}{4\theta\sum_{w=1}^d w^2\varepsilon_w}} + \zeta_n$$

for $i = 1, \dots, d-1$.

We refer the reader to Appendix A of the full version (Hahn-Klimroth and Müller, 2021) for a proof of Lemma 4.

Finally, Proposition 2 follows from an inductive argument based on the spatial coupling of the pools. More precisely, for any given compartment $V[i]$, we condition on the event that the algorithm's estimate $\tilde{\sigma}$ agrees with $(\boldsymbol{\tau}, \boldsymbol{\sigma})$ on all previous compartments $V[1], \dots, V[i-1]$. In this case, for $x \in V[i]$, $\mathcal{N}_x^j$ and $\boldsymbol{N}_x^j$ differ only in their centering, and we argue that $\mathcal{N}_x^j$ is close to $\boldsymbol{N}_x^j$, so that w.h.p., also the quantities $\mathcal{N}_x$ are concentrated well enough in between the thresholds $T^{0,1}, \dots, T^{d-1,d}$ to correctly infer all items from compartment $V[i]$. From a more quantitative point of view, a short but rather technical calculation shows that, with high probability, $\mathcal{N}_x = \boldsymbol{N}_x + \tilde{O}\left(\sqrt{s^2/\ell}\right)$, while $\boldsymbol{N}_x = \Theta(\ln s)$. By our choice of ℓ and s, the approximation error is small enough to be negligible. The proof of this assertion is carried out in detail in Appendix B of the full version (Hahn-Klimroth and Müller, 2021).

3.2. Proof of Theorem 1

Theorem 1 now is an immediate consequence of Proposition 2, if we show that Algorithm 1 computes the estimate $\tilde{\sigma}$ on the bulk items in polynomial time. To this end, it is necessary to calculate both the quantities $\mathcal{N}_x$ as well as to threshold them. However, for each item $x \in V_{\text{bulk}}$, this can be done with n^2 elementary operations. Therefore, the running time is polynomial in the number of items. ∎

4. Conclusion

On the spatially coupled pooling design For simplicity, we discuss our results for the quantitative group testing problem. Similar conclusions hold for more general $d = O(1)$. As we have seen, in the case $d = 1$, our algorithm succeeds with high probability, using a number m of tests that is within a constant factor of the information-theoretically optimal number of tests. Yet, there is a (small) gap, which can, in principle, have three reasons. Firstly, our pooling scheme might be sub-optimal. Secondly, the inference algorithm might be sub-optimal given the pooling scheme. Thirdly, there still might be a gap between efficient and non-efficient algorithms.

To discuss these three options in more detail, we briefly sketch a heuristic argument that, at least for small θ, illustrates that we cannot hope to find a better inference algorithm, given the chosen pool design. The argument resembles the information-theoretic counting bound from the introduction, and a similar line of thought was presented by Feige and Lellouche (2020). It makes use of the fact that the number of items with label 1 inside each pool is, approximately, $\mathrm{Bin}(\Gamma, k/n)$ distributed. Therefore, with high probability, all results stem from an interval of length $\ln(n)k^{1/8}$ by Chernoff's bound. To distinguish the $\binom{n}{k}$ possible values of $\boldsymbol{\sigma}$,

$$\left(\ln(n)k^{1/8}\right)^m \geq \binom{n}{k} \iff m \geq (8+o(1))\frac{1-\theta}{\theta}k$$

measurements are required asymptotically. For small θ, Algorithm 1 matches this bound. For larger values of θ, an improvement might be achievable through a tolerance for a few mistakes in the algorithm's classification, combined with a *back-propagation* approach to repair them locally. This would require less strict concentration of the number of items with specific weight in the second neighbourhood of a specific given item.

As the algorithm is optimal for small θ, the pooling design itself has to be sub-optimal. This evidence is further supported by the fact that the design would be information-theoretically optimal, if the underlying graph was much denser. More precisely, it is known that $\Gamma \sim n/2$ and $\Delta \sim m/2$ allows reconstruction of $\boldsymbol{\sigma}$ w.h.p. by exhaustive search at $m^{\mathrm{QGT}}_{\mathrm{non-ada}}$ (Gebhard et al., 2022; Scarlett and Cevher, 2017). Unfortunately, we cannot make the pooling design denser due to our inference algorithm. As the number of items with weight 1 in each compartment has to be estimated by $k\ell^{-1}$, those estimates yield approximation errors with respect to the correct underlying distribution which become more severe as the graph becomes denser. The pooling design was chosen as dense as possible such that those approximation errors are, with high probability, negligible in comparison to the difference of items with label zero and label one in the neighborhood sum. A detailed verification can be found in Appendix C of the full version (Hahn-Klimroth and Müller, 2021).

Binary group testing This paragraph discusses the similarities with as well as the differences to the inference algorithm from Coja-Oghlan et al. (2020). While the overall approach by means of a spatially coupled pooling design in combination with a thresholding algorithm has the same structure, the specific implementation is rather different.

The first difference concerns the decoding of the seed. While in Coja-Oghlan et al. (2020), the items in V_{seed} represent physical items, we introduce auxiliary items, whose labels are sampled uniformly from all assignments that posses the correct frequencies. This idea has been put forward to us by an anonymous referee. The main point here is that in the binary setting, it is possible to use simple combinatorial algorithms, like the *DD*-algorithm, on the spatially coupled pooling design in order to infer the ground-truth in V_{seed}, while using only $o(k)$ additional tests. In the quantitative

group testing setting ($d = 1$), such algorithms do also exist (Gebhard et al., 2022; Hahn-Klimroth and Kaaser, 2022). On the contrary, in the more general pooled data setting ($d > 1$), we fall short of algorithms to decode the seed on the spatially coupled design. Therefore, had we recruited the seed from V, a separate treatment of the inference of the labels in V_{seed} would have been necessary, for instance through variants of the Basis Pursuit algorithm (Donoho and Tanner, 2006).

A second, more consequential difference is our definition of weighted unexplained neighbourhood sums. More precisely, Coja-Oghlan et al. (2020) also use *unexplained neighbourhood sums* and apply a thresholding algorithm. However, the naïve combination of their spatial coupling design and thresholding based on unexplained neighbourhood sums would only lead to an improvement of a constant factor in the quantitative group testing setting. Therefore, a more substantial modification needs to be employed to actually meet the correct order of tests. The main step towards the $\ln n$ improvement is to identify the source of the failure in the above approach. This source lies in the reverse trends in the size of partial neighbourhood sums and their information content, and we overcome the ensuing effect through elaborate weighting and scaling. In the binary group testing problem, a similar trade-off can be neglected, and to the best of our knowledge, this is the first time that spatial coupling resulted in an improvement of more than a constant factor. This enhancement seems to be due to our modifications.

On multi-edges Our pool design uses multi-edges, as it simplifies the proofs. However, the same ideas should be immediately applicable to the case in which every test chooses *distinct* items. Indeed, then the unexplained neighbourhood sum follows a multivariate hypergeometric distribution (rather than being multinomially distributed). However, this distribution has similar concentration properties as the multinomial distribution.

Future directions We present an efficient algorithm that closes the previous multiplicative $\ln n$ gap between the simple information-theoretic lower bound m_{count} and all previously known efficient algorithms for the pooled data problem and its special cases. A natural question, in particular with respect to the quantitative group testing problem, is whether there is a pooling design coming with an efficient inference algorithm that matches the known exponential time constructions up to the correct constant. However, it seems unlikely that our ideas can be stretched much further onto this point. Rather, we think that one needs to come up with new ideas for a pooling scheme or a modification of the inference algorithm in order to achieve this goal.

Acknowledgments

Max Hahn-Klimroth is supported by DFG CO 646/5. Noela Müller has been supported by ERC-Grant 772606-PTRCSP. We thank Dominik Kaaser and Philipp Loick for fruitful discussions on the quantitative group testing problem. We furthermore thank an anonymous referee for his proposal of an elegant simplification of the starting phase of our algorithm.

References

N. H. Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. *Proceedings of 22nd Conference on Learning Theory (COLT)*, 2009.

E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52 (2):489–509, 2006.

A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, and P. Loick. Optimal group testing. *Proceedings of 33rd Conference on Learning Theory (COLT)*, 125:1374–1388, 2020.

A. Djackov. On a search model of false coins. *Topics in Information Theory. Hungarian Acad. Sci*, 16:163–170, 1975.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D. L. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. *2006 40th Annual Conference on Information Sciences and Systems*, pages 202–206, 2006.

A. El Alaoui, A. Ramdas, F. Krzakala, L. Zdeborová, and M. I. Jordan. Decoding from pooled data: Phase transitions of message passing. *IEEE Trans. Information Theory*, 65(1):572–585, 2019.

P. Erdős and A. Rényi. On two problems of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 8:229–243, 1963.

U. Feige and A. Lellouche. Quantitative group testing and the rank of random matrices. *arXiv:2006.09074*, 2020.

A. J. Felström and K. S. Zigangirov. Time-varying periodic convolutional codes with low-density parity-check matrix. *IEEE Transactions on Information Theory*, 45(6):2181–2191, 1999.

S. Foucart and H. Rauhut. *An Invitation to Compressive Sensing*, pages 1–39. Springer New York, New York, NY, 2013.

O. Gebhard, M. Hahn-Klimroth, D. Kaaser, and P. Loick. On the parallel reconstruction from pooled data. *Proc. 36th IEEE International Parallel & Distributed Processing Symposium (IPDPS)*, 2022.

V. Grebinski and G. Kucherov. Optimal reconstruction of graphs under the additive model. *Algorithmica*, 28(1):104–124, 2000.

M. Hahn-Klimroth and D. Kaaser. Distributed reconstruction of noisy pooled data. *Proc. 42nd IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2022.

M. Hahn-Klimroth and N. Müller. Near optimal efficient decoding from pooled data. *arXiv:2108.04342*, 2021.

E. Karimi, F. Kazemi, A. Heidarzadeh, K. R. Narayanan, and A. Sprintson. Sparse graph codes for non-adaptive quantitative group testing. *2019 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2019.

E. Karimi, F. Kazemi, A. Heidarzadeh, K. R. Narayanan, and A. Sprintson. Non-adaptive quantitative group testing using irregular sparse graph codes. *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton) IEEE*, pages 608–614, 2019.

S. Kudekar, T. Richardson, and R. Urbanke. Threshold saturation via spatial coupling: Why convolutional ldpc ensembles perform so well over the bec. *IEEE Transactions on Information Theory*, 57:803–834, 2011.

S. Kudekar, T. Richardson, and R. L. Urbanke. Spatially coupled ensembles universally achieve capacity under belief propagation. *IEEE Transactions on Information Theory*, 59(12):7761–7813, 2013.

W. Liang and J. Zou. Neural group testing to accelerate deep learning. *arXiv:2011.10704*, 2021.

J. P. Martins, R. Santos, and R. Sousa. Testing the maximum by the mean in quantitative group tests. *New Advances in Statistical Modeling and Applications*, pages 55–63, 2014.

A. Mazumdar and S. Pal. Support Recovery in Universal One-Bit Compressed Sensing. *13th Innovations in Theoretical Computer Science Conference (ITCS)*, 215:106:1–106:20, 2022.

J. Scarlett and V. Cevher. Phase transitions in the pooled data problem. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 376–384, 2017.

P. Sham, J. S. Bader, I. Craig, M. O'Donovan, and M. Owen. DNA pooling: a tool for large-scale association studies. *Nature Reviews Genetics*, 3:862–871, 2002.

H. S. Shapiro. Problem e 1399. *Amer. Math. Monthly*, 67:82, 1960.

C. Wang, Q. Zhao, and C. N. Chuah. Group testing under sum observations for heavy hitter detection. *2015 Information Theory and Applications Workshop (ITA)*, pages 149–153, 2015.

I. Wang, S. Huang, K. Lee, and K. Chen. Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms. *IEEE International Symposium on Information Theory (ISIT)*, pages 1386–1390, 2016.

L. Wang, X. Li, Y.-Q. Zhang, Y. Zhang, and K. Zhang. Evolution of scaling emergence in large-scale spatial epidemic spreading. *PLoS ONE*, 6(7):e21197, 2011.

Y. Zhou, U. Porwal, C. Zhang, H. Q. Ngo, X. Nguyen, C. Ré, and V. Govindaraju. Parallel feature selection inspired by group testing. *Advances in Neural Information Processing Systems (NeurIPS)*, 27, 2014.