Phase Transitions in the Pooled Data Problem

Jonathan Scarlett and Volkan Cevher
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
{jonathan.scarlett,volkan.cevher}@epfl.ch

Abstract

In this paper, we study the *pooled data* problem of identifying the labels associated with a large collection of items, based on a sequence of pooled tests revealing the counts of each label within the pool. In the noiseless setting, we identify an exact asymptotic threshold on the required number of tests with optimal decoding, and prove a *phase transition* between complete success and complete failure. In addition, we present a novel *noisy* variation of the problem, and provide an information-theoretic framework for characterizing the required number of tests for general random noise models. Our results reveal that noise can make the problem considerably more difficult, with strict increases in the scaling laws even at low noise levels. Finally, we demonstrate similar behavior in an *approximate recovery* setting, where a given number of errors is allowed in the decoded labels.

1 Introduction

Consider the following setting: There exists a large population of items, each of which has an associated label. The labels are initially unknown, and are to be estimated based on *pooled tests*. Each pool consists of some subset of the population, and the test outcome reveals the *total number of items* corresponding to each label that are present in the pool (but not the individual labels). This problem, which we refer to as the *pooled data* problem, was recently introduced in [1,2], and further studied in [3,4]. It is of interest in applications such as medical testing, genetics, and learning with privacy constraints, and has connections to the group testing problem [5] and its linear variants [6,7].

The best known bounds on the required number of tests under optimal decoding were given in [3]; however, the upper and lower bounds therein do not match, and can exhibit a large gap. In this paper, we completely close these gaps by providing a new lower bound that exactly matches the upper bound of [3]. These results collectively reveal a *phase transition* between success and failure, with the probability of error vanishing when the number of tests exceeds a given threshold, but tending to one below that threshold. In addition, we explore the novel aspect of random noise in the measurements, and show that this can significantly increase the required number of tests. Before summarizing these contributions in more detail, we formally introduce the problem.

1.1 Problem setup

We consider a large population of items $[p] = \{1, \dots, p\}$, each of which has an associated label in $[d] = \{1, \dots, d\}$. We let $\pi = (\pi_1, \dots, \pi_d)$ denote a vector containing the proportions of items having each label, and we assume that the vector of labels itself, $\beta = (\beta_1, \dots, \beta_p)$, is uniformly distributed over the sequences consistent with these proportions:

$$\beta \sim \mathrm{Uniform}(\mathcal{B}(\pi)), \tag{1}$$

where $\mathcal{B}(\pi)$ is the set of length-p sequences whose empirical distribution is π.

The goal is to recover β based on a sequence of pooled tests. The i-th test is represented by a (possibly random) vector $X^{(i)} \in \{0,1\}^p$, whose j-th entry $X_j^{(i)}$ indicates whether the j-th item is

Accepted to 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Sufficient for $P_{\mathrm{e}} \to 0$ [3]	Necessary for $P_{\mathrm{e}} \not\to 1$ [3]	Necessary for $P_{\mathrm{e}} \not\to 1$ (this paper)
$\frac{p}{\log p} \cdot \max_{r \in \{1,\dots,d-1\}} f(r)$	$\frac{p}{\log p} \cdot \frac{1}{2} f(1)$	$\frac{p}{\log p} \cdot \max_{r \in \{1,\dots,d-1\}} f(r)$

Table 1: Necessary and sufficient conditions on the number of tests n in the noiseless setting. The function $f(r)$ is defined in (5). Asymptotic multiplicative $1 + o(1)$ terms are omitted.

Noiseless testing	Noisy testing ($\mathrm{SNR} = p^{\Theta(1)}$)	Noisy testing ($\mathrm{SNR} = (\log p)^{\Theta(1)}$)	Noisy testing ($\mathrm{SNR} = \Theta(1)$)
$\Theta\Big(\frac{p}{\log p}\Big)$	$\Omega\Big(\frac{p}{\log p}\Big)$	$\Omega\Big(\frac{p}{\log\log p}\Big)$	$\Omega\big(p \log p\big)$

Table 2: Necessary and sufficient conditions on the number of tests n in the noisy setting. SNR denotes the signal-to-noise ratio, and the noise model is given in Section 2.2.

included in the i-th test. We define a *measurement matrix* $\mathbf{X} \in \{0,1\}^{n \times p}$ whose i-th row is given by $X^{(i)}$ for $i = 1, \dots, n$, where n denotes the total number of tests. We focus on the *non-adaptive* testing scenario, where the entire matrix $\mathbf{X}$ must be specified prior to performing any tests.

In the noiseless setting, the i-th test outcome is a vector $Y^{(i)} = (Y_1^{(i)}, \dots, Y_d^{(i)})$, with t-th entry

$$Y_t^{(i)} = N_t(\beta, X^{(i)}), \tag{2}$$

where for $t = 1, \dots, d$, we let $N_t(\beta, X) = \sum_{j \in [p]} \mathbb{1}\{\beta_j = t \cap X_j = 1\}$ denote the number of items with label t that are included in the test described by $X \in \{0,1\}^p$. More generally, in the possible presence of noise, the i-th observation is randomly generated according to

$$\big(Y^{(i)} \,|\, X^{(i)}, \beta\big) \sim P_{Y|N_1(\beta, X^{(i)}) \dots N_d(\beta, X^{(i)})} \tag{3}$$

for some conditional probability mass function $P_{Y|N_1,\dots,N_d}$ (or density function in the case of continuous observations). We assume that the observations $Y^{(i)}$ ($i = 1, \dots, n$) are conditionally independent given $\mathbf{X}$, but otherwise make no assumptions on $P_{Y|N_1,\dots,N_d}$. Clearly, the noiseless model (2) falls under this more general setup.

Similarly to $\mathbf{X}$, we let $\mathbf{Y}$ denote an $n \times d$ matrix of observations, with the i-th row being $Y^{(i)}$. Given $\mathbf{X}$ and $\mathbf{Y}$, a *decoder* outputs an estimate $\hat{\beta}$ of β, and the error probability is given by

$$P_{\mathrm{e}} = \mathbb{P}[\hat{\beta} \neq \beta], \tag{4}$$

where the probability is with respect to β, $\mathbf{X}$, and $\mathbf{Y}$. We seek to find conditions on the number of tests n under which P_{e} attains a certain target value in the limit as $p \to \infty$, and our main results provide necessary conditions (i.e., lower bounds on n) for this to occur. As in [3], we focus on the case that d and π are fixed and do not depend on p.[1]

1.2 Contributions and comparisons to existing bounds

Our focus in this paper is on *information-theoretic* bounds on the required number of tests that hold regardless of practical considerations such as computation and storage. Among the existing works in the literature, the one most relevant to this paper is [3], whose bounds strictly improve on the initial bounds in [1]. The same authors also proved a phase transition for a *practical* algorithm based on approximate message passing [4], but the required number of tests is in fact significantly larger than the information-theoretic threshold (specifically, linear in p instead of sub-linear).

Table 1 gives a summary of the bounds from [3] and our contributions in the noiseless setting. To define the function $f(r)$ therein, we introduce the additional notation that for $r = \{1, \dots, d-1\}$, $\pi^{(r)} = (\pi_1^{(r)}, \dots, \pi_r^{(r)})$ is a vector whose first entry sums the largest $d - r + 1$ entries of π, and whose remaining entries coincide with the remaining $r - 1$ entries of π. We have

$$f(r) = \max_{r \in \{1,\dots,d-1\}} \frac{2(H(\pi) - H(\pi^{(r)}))}{d - r}, \tag{5}$$

meaning that the entries in Table 1 corresponding to the results of [3] are given as follows:

[1] More precisely, π should be rounded to the nearest empirical distribution (e.g., in ℓ_∞-norm) for sequences $\beta \in [d]^p$ of length p; we leave such rounding implicit throughout the paper.

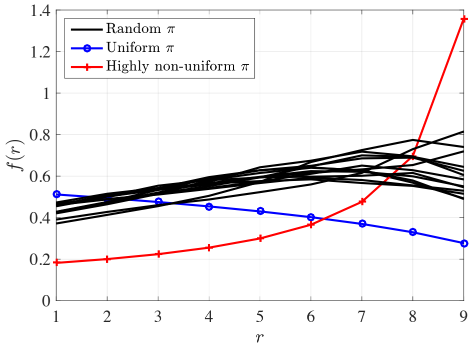

Figure 1: The function $f(r)$ in (5), for several choices of π, with $d = 10$. The random π are drawn uniformly on the probability simplex, and the highly non-uniform choice of π is given by $\pi = (0.49, 0.49, 0.0025, \dots, 0.0025)$. When the maximum is achieved at $r = 1$, the bounds of [3] coincide up to a factor of two, whereas if the maximum is achieved for $r > 1$ then the gap is larger.

- (Achievability) When the entries of $\mathbf{X}$ are i.i.d. on $\mathrm{Bernoulli}(q)$ for some $q \in (0,1)$ (not depending on p), there exists a decoder such that $P_{\mathrm{e}} \to 0$ as $p \to \infty$ with

$$n \le \frac{p}{\log p}\left(\max_{r\in\{1,\dots,d-1\}} \frac{2(H(\pi) - H(\pi^{(r)}))}{d-r}\right)(1+\eta) \tag{6}$$

  for arbitrarily small $\eta > 0$.

- (Converse) In order to achieve $P_{\mathrm{e}} \not\to 1$ as $p \to \infty$, it is necessary that

$$n \ge \frac{p}{\log p}\left(\frac{H(\pi)}{d-1}\right)(1-\eta) \tag{7}$$

  for arbitrarily small $\eta > 0$.

Unfortunately, these bounds do not coincide. If the maximum in (6) is achieved by $r = 1$ (which occurs, for example, when π is uniform [3]), then the gap only amounts to a factor of two. However, as we show in Figure 1, if we compute the bounds for some "random" choices of π then the gap is typically larger (i.e., $r = 1$ does not achieve the maximum), and we can construct choices where the gap is significantly larger. Closing these gaps was posed as a key open problem in [3].

We can now summarize our contributions as follows:

1. We give a lower bound that exactly matches (6), thus completely closing the above-mentioned gaps in the existing bounds and solving the open problem raised in [3]. More specifically, we show that $P_{\mathrm{e}} \to 1$ whenever $n \le \frac{p}{\log p}\big(\max_{r\in\{1,\dots,d-1\}} \frac{2(H(\pi)-H(\pi^{(r)}))}{d-r}\big)(1-\eta)$ for some $\eta > 0$, thus identifying an exact *phase transition* – a threshold above which the error probability vanishes, but below which the error probability tends to one.
2. We develop a framework for understanding variations of the problem consisting of random noise, and give an example of a noise model where the scaling laws are strictly higher compared to the noiseless case. A summary is given in Table 2; the case $\mathrm{SNR} = (\log p)^{\Theta(1)}$ reveals a strict increase in the scaling laws even when the signal-to-noise ratio grows unbounded, and the case $\mathrm{SNR} = \Theta(1)$ reveals that the required number of tests increases from sub-linear to super-linear in the dimension when the signal-to-noise ratio is constant.
3. In the supplementary material, we discuss how our lower bounds extend readily to the *approximate recovery* criterion, where we only require β to be identified up to a certain Hamming distance. However, for clarity, we focus on exact recovery throughout the paper.

In a recent independent work [8], an *adversarial* noise setting was introduced. This turns out to be fundamentally different to our noisy setting. In particular, the results of [8] state that exact recovery is impossible, and even with approximate recovery, a huge number of tests (i.e., higher than polynomial) is needed unless $\Delta = O\big(q_{\max}^{1/2+o(1)}\big)$, where $q_{\max}$ is the maximum allowed reconstruction error measured by the Hamming distance, and Δ is maximum adversarial noise amplitude. Of course, both random and adversarial noise are of significant interest, depending on the application.

Notation. For a positive integer d, we write $[d] = \{1, \dots, d\}$. We use standard information-theoretic notations for the (conditional) entropy and mutual information, e.g., $H(X)$, $H(Y|X)$, $I(X;Y|Z)$ [9]. All logarithms have base e, and accordingly, all of the preceding information measures are in units of nats. The Gaussian distribution with mean μ and variance σ^2 is denoted by $\mathrm{N}(\mu, \sigma^2)$. We use the standard asymptotic notations $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, $\omega(\cdot)$ and $\Theta(\cdot)$.

2 Main results

In this section, we present our main results for the noiseless and noisy settings. The proofs are given in Section 3, as well as the supplementary material.

2.1 Phase transition in the noiseless setting

The following theorem proves that the upper bound given in (6) is tight. Recall that for $r = \{1, \dots, d-1\}$, $\pi^{(r)} = (\pi_1^{(r)}, \dots, \pi_r^{(r)})$ is a vector whose first entry sums the largest $d - r + 1$ entries of π, and whose remaining entries coincide with the remaining $r - 1$ entries of π.

Theorem 1. (Noiseless setting) *Consider the pooled data problem described in Section 1.1 with a given number of labels d and label proportion vector π (not depending on the dimension p). For any decoder, in order to achieve $P_{\mathrm{e}} \not\to 1$ as $p \to \infty$, it is necessary that*

$$n \geq \frac{p}{\log p} \left(\max_{r \in \{1, \dots, d-1\}} \frac{2(H(\pi) - H(\pi^{(r)}))}{d - r} \right) (1 - \eta) \tag{8}$$

for arbitrarily small $\eta > 0$.

Combined with (6), this result reveals an exact *phase transition* on the required number of measurements: Denoting $n^* = \frac{p}{\log p} \big(\max_{r \in \{1, \dots, d-1\}} \frac{2(H(\pi) - H(\pi^r))}{d - r} \big)$, the error probability vanishes for $n \geq n^*(1 + \eta)$, tends to one for $n \leq n^*(1 - \eta)$, regardless of how small η is chosen to be.

Remark 1. Our model assumes that β is uniformly distributed over the sequences with empirical distribution π, whereas [3] assumes that β is i.i.d. on π. However, Theorem 1 readily extends to the latter setting: Under the i.i.d. model, once we condition on a given empirical distribution, the conditional distribution of β is uniform. As a result, the converse bound for the i.i.d. model follows directly from Theorem 1 by basic concentration and the continuity of the entropy function.

2.2 Information-theoretic framework for the noisy setting

We now turn to *general noise models* of the form (3), and provide necessary conditions for the noisy pooled data problem in terms of the mutual information. General characterizations of this form were provided previously for group testing [10, 11] and other sparse recovery problems [12, 13].

Our general result is stated in terms of a maximization over a vector parameter $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d)$ with $\ell_t \in \{0, \dots, \pi_t p\}$ for all t. We will see in the proof that ℓ_t represents the number of items of type t that are unknown to the decoder after $p\pi_t - \ell_t$ are revealed by a genie. We define the following:

- Given $\boldsymbol{\ell}$ and β, we let $S_{\boldsymbol{\ell}}$ be a random set of indices in $[p]$ such that for each $t \in [d]$, the set contains ℓ_t indices corresponding to entries where β equals t. Specifically, we define $S_{\boldsymbol{\ell}}$ to be uniformly distributed over all such sets. Moreover, we define $S_{\boldsymbol{\ell}}^c = [p] \setminus S_{\boldsymbol{\ell}}$.
- Given the above definitions, we define

  $$\beta_{S_{\boldsymbol{\ell}}^c} = \begin{cases} \beta_j & j \in S_{\boldsymbol{\ell}}^c \\ \star & \text{otherwise}, \end{cases} \tag{9}$$

  where $\star$ can be thought of as representing an unknown value. Hence, knowing $\beta_{S_{\boldsymbol{\ell}}^c}$ amounts to knowing the labels of all items in the set $S_{\boldsymbol{\ell}}^c$.
- We define $|\mathcal{B}_{\boldsymbol{\ell}}(\pi)|$ to be the number of sequences $\beta \in [d]^p$ that coincide with a given $\beta_{S_{\boldsymbol{\ell}}^c}$ on the entries not equaling $\star$, while also having empirical distribution π overall. This number does not depend on the specific choice of $S_{\boldsymbol{\ell}}^c$. As an example, when $\ell_t = p\pi_t$ for all t, we have $S_{\boldsymbol{\ell}} = [p]$, $\beta_{S_{\boldsymbol{\ell}}^c} = (\star, \dots, \star)$, and $|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| = |\mathcal{B}(\pi)|$, defined following (1)
- We let $\|\boldsymbol{\ell}\|_0$ denote the number of values in $(\ell_1, \dots, \ell_d)$ that are positive.

With these definitions, we have the following result for general random noise models.

Theorem 2. (Noisy setting) *Consider the pooled data problem described in Section 1.1 under a general observation model of the form* (3)*, with a given number of labels* d *and label proportion vector* π*. For any decoder, in order to achieve* $P_{\rm e} \le \delta$ *for a given* $\delta \in (0,1)$*, it is necessary that*

$$n \ge \max_{\boldsymbol{\ell} \,:\, \|\boldsymbol{\ell}\|_0 \ge 2} \frac{\big(\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)|\big)(1-\delta) - \log 2}{\frac{1}{n}\sum_{i=1}^n I(\beta; Y^{(i)} | \beta_{S_{\boldsymbol{\ell}}^c}, X^{(i)})}. \tag{10}$$

In order to obtain more explicit bounds on n from (10), one needs to characterize the mutual information terms, ideally forming an upper bound that does not depend on the distribution of the measurement matrix $\mathbf{X}$. We do this for some specific models below; however, in general it can be a difficult task. The following corollary reveals that if the entries of $\mathbf{X}$ are i.i.d. on $\mathrm{Bernoulli}(q)$ for some $q \in (0,1)$ (as was assumed in [3]), then we can simplify the bound.

Corollary 1. (Noisy setting with Bernoulli testing) *Suppose that the entries of* $\mathbf{X}$ *are i.i.d. on* $\mathrm{Bernoulli}(q)$ *for some* $q \in (0,1)$*. Under the setup of Theorem 2, it is necessary that*

$$n \ge \max_{\boldsymbol{\ell} \,:\, \|\boldsymbol{\ell}\|_0 \ge 2} \frac{\big(\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)|\big)(1-\delta) - \log 2}{I(X_{0,\boldsymbol{\ell}}; Y | X_{1,\boldsymbol{\ell}})}, \tag{11}$$

where $(X_{0,\boldsymbol{\ell}}, X_{1,\boldsymbol{\ell}}, Y)$ *are distributed as follows: (i)* $X_{0,\boldsymbol{\ell}}$ *(respectively,* $X_{1,\boldsymbol{\ell}}$*) is a concatenation of the vectors* $X_{0,\boldsymbol{\ell}}(1), \dots, X_{0,\boldsymbol{\ell}}(d)$ *(respectively,* $X_{1,\boldsymbol{\ell}}(1), \dots, X_{1,\boldsymbol{\ell}}(d)$*), the* t*-th of which contains* ℓ_t *(respectively,* $\pi_t p - \ell_t$*) entries independently drawn from* $\mathrm{Bernoulli}(q)$*; (ii) Letting each* N_t *(*$t = 1, \dots, d$*) be the total number of ones in* $X_{0,\boldsymbol{\ell}}(t)$ *and* $X_{1,\boldsymbol{\ell}}(t)$ *combined, the random variable* Y *is drawn from* $P_{Y|N_1,\dots,N_d}$ *according to* (3).

As well as being simpler to evaluate, this corollary may be of interest in scenarios where one does not have complete freedom in designing $\mathbf{X}$, and one instead insists on using Bernoulli testing. For instance, one may not know how to optimize $\mathbf{X}$, and accordingly resort to generating it at random.

Example 1: Application to the noiseless setting. In the supplementary material, we show that in the noiseless setting, Theorem 2 recovers a weakened version of Theorem 1 with $1-\eta$ replaced by $1-\delta-o(1)$ in (8). Hence, while Theorem 2 does not establish a phase transition, it does recover the exact threshold on the number of measurements required to obtain $P_{\rm e} \to 0$.

An overview of the proof of this claim is as follows. We restrict the maximum in (10) to choices of $\boldsymbol{\ell}$ where each ℓ_t equals either its minimum value 0 or its maximum value $p\pi_t$. Since we are in the noiseless setting, each mutual information term reduces to the conditional entropy of $Y^{(i)} = (Y_1^{(i)}, \dots, Y_d^{(i)})$ given $\beta_{S_{\boldsymbol{\ell}}^c}$ and $X^{(i)}$. For the values of t such that $\ell_t = 0$, the value $Y_t^{(i)}$ is deterministic (i.e., it has zero entropy), whereas for the values of t such that $\ell_t = p\pi_t$, the value $Y_t^{(i)}$ follows a hypergeometric distribution, whose entropy behaves as $\big(\frac{1}{2}\log p\big)(1+o(1))$.

In the case that $\mathbf{X}$ is i.i.d. on $\mathrm{Bernoulli}(q)$, we can use Corollary 1 to obtain the following necessary condition for $P_{\rm e} \le \delta$ as as $p \to \infty$, proved in the supplementary material:

$$n \ge \frac{p}{\log(pq(1-q))}\bigg(\max_{r \in \{1,\dots,d-1\}} \frac{2(H(\pi) - H(\pi^r))}{d-r}\bigg)(1-\delta-o(1)) \tag{12}$$

for any $q = q(p)$ such that both q and $1-q$ behave as $\omega\big(\frac{1}{p}\big)$. Hence, while $q = \Theta(1)$ recovers the threshold in (8), the required number of tests strictly increases when $q = o(1)$, albeit with a mild logarithmic dependence.

Example 2: Group testing. To highlight the versatility of Theorem 2 and Corollary 1, we show that the latter recovers the lower bounds given in the group testing framework of [11].

Set $d = 2$, and let label 1 represent "defective" items, and label 2 represent "non-defective" items. Let $P_{Y|N_1N_2}$ be of the form $P_{Y|N_1}$ with $Y \in \{0,1\}$, meaning the observations are binary and depend only on the number of defective items in the test. For brevity, let $k = p\pi_1$ denote the total number of defective items, so that $p\pi_2 = p - k$ is the number of non-defective items.

Letting $\ell_2 = p - k$ in (11), and letting ℓ_1 remain arbitrary, we obtain the necessary condition

$$n \ge \max_{\ell_1 \in \{1,\dots,k\}} \frac{\big(\log\binom{p-k+\ell_1}{\ell_1}\big)(1-\delta) - \log 2}{I(X_{0,\ell_1}; Y | X_{1,\ell_1})}, \tag{13}$$

where X_{0,ℓ_1} is a shorthand for $X_{0,\boldsymbol{\ell}}$ with $\boldsymbol{\ell} = (\ell_1, p-k)$, and similarly for X_{1,ℓ_1}. This matches the lower bound given in [11] for Bernoulli testing with general noise models, for which several corollaries for specific models were also given.

Example 3: Gaussian noise. To give a concrete example of a noisy setting, consider the case that we observe the values in (2), but with each such value corrupted by independent Gaussian noise:

$$Y_t^{(i)} = N_t(\beta, X^{(i)}) + Z_t^{(i)}, \tag{14}$$

where $Z_t^{(i)} \sim \mathrm{N}(0, p\sigma^2)$ for some $\sigma^2 > 0$. Note that given $X^{(i)}$, the values N_t themselves have variance at most proportional to p (e.g., see Appendix C), so $\sigma^2 = \Theta(1)$ can be thought of as the constant signal-to-noise ratio (SNR) regime.

In the supplementary material, we prove the following bounds for this model:

- By letting each ℓ_t in (10) equal its minimum or maximum value analogously to the noiseless case above, we obtain the following necessary condition for $P_{\mathrm{e}} \le \delta$ as $p \to \infty$:

$$n \ge \bigg(\max_{G \subseteq [d] \,:\, |G| \ge 2} \frac{p_G H(\pi_G)}{\sum_{t \in G} \frac{1}{2} \log\big(1 + \frac{\pi_t}{4\sigma^2}\big)} \bigg)(1 - \delta - o(1)), \tag{15}$$

  where $p_G := \sum_{t \in G} \pi_t p$, and π_G has entries $\frac{\pi_t}{\sum_{t' \in G} \pi_{t'}}$ for $t \in G$. Hence, we have the following:

  - In the case that $\sigma^2 = p^{-c}$ for some $c \in (0,1)$, each summand in the denominator simplifies to $\big(\frac{c}{2} \log p\big)(1 + o(1))$, and we deduce that compared to the noiseless case (*cf.*, (8)), the asymptotic number of tests increases by at least a constant factor of $\frac{1}{c}$.
  - In the case that $\sigma^2 = (\log p)^{-c}$ for some $c > 0$, each summand in the denominator simplifies to $\big(\frac{c}{2} \log\log p\big)(1 + o(1))$, and we deduce that compared to the noiseless case, the asymptotic number of tests increases by at least a factor of $\frac{\log p}{c \log\log p}$. Hence, we observe a strict increase in the scaling laws despite the fact that the SNR grows unbounded.
  - While (15) also provides an $\Omega(p)$ lower bound for the case $\sigma^2 = \Theta(1)$, we can in fact do better via a different choice of $\boldsymbol{\ell}$ (see below).

- By letting $\ell_1 = p\pi_1$, $\ell_2 = 1$, and $\ell_t = 0$ for $t = 3, \dots, d$, we obtain the necessary condition

$$n \ge \big(4p\sigma^2 \log p\big)(1 - \delta - o(1)) \tag{16}$$

  for $P_{\mathrm{e}} \le \delta$ as $p \to \infty$. Hence, if $\sigma^2 = \Theta(1)$, we require $n = \Omega(p \log p)$; this is super-linear in the dimension, in contrast with the sub-linear $\Theta\big(\frac{p}{\log p}\big)$ behavior observed in the noiseless case. Note that this choice of $\boldsymbol{\ell}$ essentially captures the difficulty in identifying a *single* item, namely, the one corresponding to $\ell_2 = 1$.

These findings are summarized in Table 2; see also the supplementary material for extensions to the approximate recovery setting.

Remark 2. While it may seem unusual to add continuous noise to discrete observations, this still captures the essence of the noisy pooled data problem, and simplifies the evaluation of the mutual information terms in (10). Moreover, this converse bound immediately implies the same bound for the *discrete* model in which the noise consists of adding a Gaussian term, rounding, and clipping to $\{0, \dots, p\}$, since the decoder could always choose to perform these operations as pre-processing.

3 Proofs

Here we provide the proof of Theorem 1, along with an overview of the proof of Theorem 2. The remaining proofs are given in the supplementary material.

3.1 Proof of Theorem 1

Step 1: Counting typical outcomes. We claim that it suffices to consider the case that $\mathbf{X}$ is deterministic and $\hat{\beta}$ is a deterministic function of $\mathbf{Y}$; to see this, we note that when either of these are random we have $P_{\mathrm{e}} = \mathbb{E}_{\mathbf{X},\hat{\beta}}[\mathbb{P}_\beta[\mathrm{error}]]$, and the average is lower bounded by the minimum.

The following lemma, proved in the supplementary material, shows that for any $X^{(i)}$, each entry of the corresponding outcome $Y^{(i)}$ lies in an interval of length $O\big(\sqrt{p \log p}\big)$ with high probability.

Lemma 1. *For any deterministic test vector $X \in \{0,1\}^p$, and for β uniformly distributed on $\mathcal{B}(\pi)$, we have for each $t \in [d]$ that*

$$\mathbb{P}\Big[\big|N_t(\beta, X) - \mathbb{E}[N_t(\beta, X)]\big| > \sqrt{p\log p}\Big] \le \frac{2}{p^2}. \tag{17}$$

By Lemma 1 and the union bound, we have with probability at least $1 - \frac{2nd}{p^2}$ that $\big|N_t(\beta, X^{(i)}) - \mathbb{E}[N_t(\beta, X^{(i)})]\big| \le \sqrt{p\log p}$ for all $i \in [n]$ and $t \in [d]$. Letting this event be denoted by $\mathcal{A}$, we have

$$P_{\rm e} \ge \mathbb{P}[\mathcal{A}] - \mathbb{P}[\mathcal{A} \cap \text{no error}] \ge 1 - \frac{2nd}{p^2} - \mathbb{P}[\mathcal{A} \cap \text{no error}]. \tag{18}$$

Next, letting $\mathbf{Y}(\beta) \in [p]^{n\times d}$ denote $\mathbf{Y}$ explicitly as a function of β and similarly for $\hat{\beta}(\mathbf{Y}) \in [d]^p$, and letting $\mathcal{Y}_{\mathcal{A}}$ denote the set of matrices $\mathbf{Y}$ under which the event $\mathcal{A}$ occurs, we have

$$\mathbb{P}[\mathcal{A} \cap \text{no error}] = \frac{1}{|\mathcal{B}(\pi)|} \sum_{b \in \mathcal{B}(\pi)} \mathbb{1}\{\mathbf{Y}(b) \in \mathcal{Y}_{\mathcal{A}} \cap \hat{\beta}(\mathbf{Y}(b)) = b\} \tag{19}$$

$$\le \frac{|\mathcal{Y}_{\mathcal{A}}|}{|\mathcal{B}(\pi)|}, \tag{20}$$

where (20) follows since each each $\mathbf{Y} \in \mathcal{Y}_{\mathcal{A}}$ can only be counted once in the summation of (19), due to the condition $\hat{\beta}(\mathbf{Y}(b)) = b$.

Step 2: Bounding the set cardinalities. By a standard combinatorial argument (e.g., [14, Ch. 2]) and the fact that π is fixed as $p \to \infty$, we have

$$|\mathcal{B}(\pi)| = e^{p(H(\pi)+o(1))}. \tag{21}$$

To bound $|\mathcal{Y}_{\mathcal{A}}|$, first note that the entries of each $Y^{(i)} \in [p]^d$ sum to a deterministic value, namely, the number of ones in $X^{(i)}$. Hence, each $\mathbf{Y} \in \mathcal{Y}_{\mathcal{A}}$ is uniquely described by a sub-matrix of $\mathbf{Y} \in [p]^{n\times d}$ of size $n \times (d-1)$. Moreover, since $\mathcal{Y}_{\mathcal{A}}$ only includes matrices under which $\mathcal{A}$ occurs, each value in this sub-matrix only takes one of at most $2\sqrt{p\log p} + 1$ values. As a result, we have

$$|\mathcal{Y}_{\mathcal{A}}| \le \big(2\sqrt{p\log p} + 1\big)^{n(d-1)}, \tag{22}$$

and combining (18)–(22) gives

$$P_{\rm e} \ge \frac{\big(2\sqrt{p\log p} + 1\big)^{n(d-1)}}{e^{p(H(\pi)+o(1))}} - \frac{2nd}{p^2}. \tag{23}$$

Since d is constant, it immediately follows that $P_{\rm e} \to 1$ whenever $n \le \frac{pH(\pi)}{(d-1)\log(2\sqrt{p\log p}+1)}(1-\eta)$ for some $\eta > 0$. Applying $\log(2\sqrt{p\log p} + 1) = \big(\frac{1}{2}\log p\big)(1 + o(1))$, we obtain the following necessary condition for $P_{\rm e} \not\to 1$:

$$n \ge \frac{2pH(\pi)}{(d-1)\log p}(1-\eta). \tag{24}$$

This yields the term in (8) corresponding to $r = 1$.

Step 3: Genie argument. Let G be a subset of $[d]$ of cardinality at least two, and define $G^c = [d]\backslash G$. Moreover, define β_{G^c} to be a length-p vector with

$$(\beta_{G^c})_j = \begin{cases} \beta_j & \beta_j \in G^c \\ \star & \beta_j \in G, \end{cases} \tag{25}$$

where the symbol $\star$ can be thought of as representing an unknown value. We consider a modified setting in which a genie reveals β_{G^c} to the decoder, i.e., the decoder knows the labels of all items for which the label lies in G^c, and is only left to estimate those in G. This additional knowledge can only make the pooled data problem easier, and hence, any lower bound in this modified setting remains valid in the original setting.

In the genie-aided setting, instead of receiving the full observation vector $Y^{(i)} = (Y_1^{(i)}, \dots, Y_d^{(i)})$, it is equivalent to only be given $\{Y_j^{(i)} : j \in G\}$, since the values in G^c are uniquely determined from β_{G^c} and $X^{(i)}$. This means that the genie-aided setting can be cast in the original setting with modified parameters: (i) p is replaced by $p_G = \sum_{t \in G} \pi_t p$, the number of items with unknown labels; (ii) d is replaced by $|G|$, the number of distinct remaining labels; (iii) π is replaced by π_G, defined to be a $|G|$-dimensional probability vector with entries equaling $\frac{\pi_t}{\sum_{t' \in G} \pi_{t'}}$ $(t \in G)$.

Due to this equivalence, the condition (24) yields the necessary condition $n \geq \frac{2 p_G H(\pi_G)}{(|G|-1)\log p}(1-\eta)$, and maximizing over all G with $|G| \geq 2$ gives

$$n \geq \max_{G \subseteq [d] \,:\, |G| \geq 2} \frac{2 p_G H(\pi_G)}{(|G|-1)\log p}\big(1-\eta\big). \tag{26}$$

Step 4: Simplification. Define $r = d - |G| + 1$. We restrict the maximum in (26) to sets G indexing the highest $|G| = d - r + 1$ values of π, and consider the following process for sampling from π:

- Draw a sample v from $\pi^{(r)}$ (defined above Theorem 1);
- If v corresponds to the first entry of $\pi^{(r)}$, then draw a random sample from π_G and output it as a label (i.e., the labels have conditional probability proportional to the top $|G|$ entries of π);
- Otherwise, if v corresponds to one of the other entries of $\pi^{(r)}$, then output v as a label.

By Shannon's property of entropy for sequentially-generated random variables [15, p. 10], we find that $H(\pi) = H(\pi^{(r)}) + \big(\sum_{t \in G} \pi_t\big) H(\pi_G)$. Moreover, since $p_G = p \cdot \sum_{t \in G} \pi_j$, this can be written as $p_G H(\pi_G) = p\big(H(\pi) - H(\pi^{(r)})\big)$. Substituting into (26), noting that $|G| - 1 = d - r$ by the definition of r, and maximizing over $r = 1, \dots, d-1$, we obtain the desired result (8).

3.2 Overview of proof of Theorem 2

We can interpret the pooled data problem as a communication problem in which a "message" β is sent over a "channel" $P_{Y|N_1,\dots,N_d}$ via "codewords" of the form $\{(N_1^{(i)}, \dots, N_d^{(i)})\}_{i=1}^n$ that are constructed by summing various columns of $\mathbf{X}$. As a result, it is natural to use Fano's inequality [9, Ch. 7] to lower bound the error probability in terms of information content (entropy) of β and the amount of information that $\mathbf{Y}$ reveals about β (mutual information).

However, a naive application of Fano's inequality only recovers the bound in (10) with $\boldsymbol{\ell} = p\pi$. To handle the other possible choices of $\boldsymbol{\ell}$, we again consider a *genie-aided setting* in which, for each $t \in [d]$, the decoder is informed of $p\pi_t - \ell_t$ of the items whose label equals t. Hence, it only remains to identify the remaining ℓ_t items of each type. This genie argument is a generalization of that used in the proof of Theorem 1, in which each ℓ_t was either equal to its minimum value zero or its maximum value $p\pi_t$. In Example 3 of Section 2, we saw that this generalization can lead to a strictly better lower bound in certain noisy scenarios.

The complete proof of Theorem 2 is given in the supplementary material.

4 Conclusion

We have provided novel information-theoretic lower bounds for the pooled data problem. In the noiseless setting, we provided a matching lower bound to the upper bound of [3], establishing an exact threshold indicating a phase transition between success and failure. In the noisy setting, we provided a characterization of general noise models in terms of the mutual information. In the special case of Gaussian noise, we proved an inherent added difficulty compared to the noiseless setting, with strict increases in the scaling laws even when the signal-to-noise ratio grows unbounded.

An interesting direction for future research is to provide *upper bounds* for the noisy setting, potentially establishing the tightness of Theorem 2 for general noise models. This appears to be challenging using existing techniques; for instance, the pooled data problem bears similarity to group testing with *linear* sparsity, whereas existing mutual information based upper bounds for group testing are limited to the *sub-linear* regime [10, 11, 16]. In particular, the proofs of such bounds are based on concentration inequalities which, when applied to the linear regime, lead to additional requirements on the number of tests that prevent tight performance characterizations.

Acknowledgment: This work was supported in part by the European Commission under Grant ERC Future Proof, SNF Sinergia project CRSII2-147633, SNF 200021-146750, and EPFL Fellows Horizon2020 grant 665667.

References

[1] I.-H. Wang, S. L. Huang, K. Y. Lee, and K. C. Chen, “Data extraction via histogram and arithmetic mean queries: Fundamental limits and algorithms,” in *IEEE Int. Symp. Inf. Theory*, July 2016, pp. 1386–1390.

[2] I.-H. Wang, S. L. Huang, and K. Y. Lee, “Extracting sparse data via histogram queries,” in *Allerton Conf. Comm., Control, and Comp.*, 2016.

[3] A. E. Alaoui, A. Ramdas, F. Krzakala, L. Zdeborova, and M. I. Jordan, “Decoding from pooled data: Sharp information-theoretic bounds,” 2016, http://arxiv.org/abs/1611.09981.

[4] ——, “Decoding from pooled data: Phase transitions of message passing,” 2017, http://arxiv.org/abs/1702.02279.

[5] D.-Z. Du and F. K. Hwang, *Combinatorial group testing and its applications*, ser. Series on Applied Mathematics. World Scientific, 1993.

[6] A. Sebő, “On two random search problems,” *J. Stat. Plan. Inf.*, vol. 11, no. 1, pp. 23–31, 1985.

[7] M. Malyutov and H. Sadaka, “Maximization of ESI. Jaynes principle for testing significant inputs of linear model,” *Rand. Opt. Stoch. Eq.*, vol. 6, no. 4, pp. 339–358, 1998.

[8] W.-N. Chen and I.-H. Wang, “Partial data extraction via noisy histogram queries: Information theoretic bounds,” in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2017.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[10] M. Malyutov, “The separating property of random matrices,” *Math. Notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.

[11] G. Atia and V. Saligrama, “Boolean compressed sensing and noisy group testing,” *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, March 2012.

[12] C. Aksoylar, G. K. Atia, and V. Saligrama, “Sparse signal processing with linear and nonlinear observations: A unified Shannon-theoretic approach,” *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 749–776, Feb. 2017.

[13] J. Scarlett and V. Cevher, “Limits on support recovery with probabilistic models: An information-theoretic framework,” *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 593–620, 2017.

[14] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.

[15] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. Journal*, vol. 27, pp. 379–423, July and Oct. 1948.

[16] J. Scarlett and V. Cevher, “Phase transitions in group testing,” in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[17] W. Hoeffding, “Probability inequalities for sums of bounded random variables,” *J. Amer. Stat. Assoc.*, vol. 58, no. 301, pp. 13–30, 1963.

[18] J. Massey, “On the entropy of integer-valued random variables,” in *Int. Workshop on Inf. Theory*, 1988.

[19] G. Reeves and M. Gastpar, “The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3065–3092, May 2012.

[20] ——, “Approximate sparsity pattern recovery: Information-theoretic lower bounds,” *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3451–3465, June 2013.

[21] J. Scarlett and V. Cevher, “How little does non-exact recovery help in group tesitng?” in *IEEE Int. Conf. Acoust. Sp. Sig. Proc. (ICASSP)*, New Orleans, 2017.

[22] ——, “On the difficulty of selecting Ising models with approximate recovery,” *IEEE Trans. Sig. Inf. Proc. over Networks*, vol. 2, no. 4, pp. 625–638, 2016.

[23] J. C. Duchi and M. J. Wainwright, “Distance-based and continuum Fano inequalities with applications to statistical estimation,” 2013, http://arxiv.org/abs/1311.2669.

Supplementary Material

"Phase Transitions in the Pooled Data Problem"

(Jonathan Scarlett and Volkan Cevher, NIPS 2017)

Note that the references for the citations below are given in the main document.

A Proof of Lemma 1

Let N_t be a shorthand for $N_t(\beta, X)$. Since β uniformly distributed on the set of sequences with empirical distribution π, and N_t counts the number of locations where $X_j = 1$ and $\beta_j = t$, we have $N_t \sim \text{Hypergeometric}(\pi_t p, m(X), p)$, where $m(X)$ denotes the number of ones in X, and $\text{Hypergeometric}(k, m, p)$ denotes the distribution of the number of "special items" when k items are drawn from a population of p items (m of which are special) without replacement.

As a result, by Hoeffding's inequality for sampling without replacement [17], we have $\mathbb{P}[|N_1 - \mathbb{E}[N_1]| > \delta p] \le 2e^{-2\delta^2 p}$. Choosing $\delta = \sqrt{\frac{\log p}{p}}$ yields (17).

B Proofs of general results for the noisy setting

Throughout this section, the random variables $\mathbf{X}$ and β are always discrete, whereas we allow the observations $\mathbf{Y}$ to be either discrete or continuous. In the continuous case, entropy terms should be interpreted as being the *differential entropy* [9, Ch. 8].

B.1 Proof of Theorem 2

Throughout the proof, we make use of the definitions in Section 2.2 in terms of a given vector of integers $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d)$ with $0 \le \ell_t \le \pi_t p$. Note that we only consider choices of $\boldsymbol{\ell}$ such that $\|\boldsymbol{\ell}\|_0 \ge 2$, since otherwise the recovery problem would be trivial (e.g., if only a single ℓ_t is positive, then one achieves zero error probability be estimating all unknown labels to be t).

Step 1: Fano's inequality and a genie argument. As outlined in Section 3, a natural starting point is to apply *Fano's inequality* [9, Sec. 2.10] to obtain

$$I(\beta; \mathbf{Y}|\mathbf{X}) \ge \log|\mathcal{B}(\pi)| \cdot (1-\delta) - \log 2. \tag{27}$$

Unfortunately, this bound alone is not sufficient to attain the desired result. To do that, we apply a *genie argument*, considering the following modified setting:

- The items $[p]$ are split into $S_{\boldsymbol{\ell}}$ (*cf.*, Section 2.2) and $S_{\boldsymbol{\ell}}^c = [p] \backslash S_{\boldsymbol{\ell}}$;
- A genie reveals to the decoder the labels of all items in $S_{\boldsymbol{\ell}}^c$, or equivalently, the vector $\beta_{S_{\boldsymbol{\ell}}^c}$ defined in (9);
- The decoder is left to identify only the entries in β indexed by $S_{\boldsymbol{\ell}}$, i.e., to "fill in" the indices of $\beta_{S_{\boldsymbol{\ell}}^c}$ that are equal to $\star$.

Clearly the additional information at the decoder only makes the recovery problem easier, and thus any lower bound for the genie-aided setting is also a lower bound for the original setting.

Let us condition on particular realizations of $\beta_{S_{\boldsymbol{\ell}}^c} = b_{S_{\boldsymbol{\ell}}^c}$, and $\mathbf{X} = \mathbf{x}$, and let $\delta(b_{S_{\boldsymbol{\ell}}^c}, \mathbf{x})$ denote the corresponding conditional error probability. For any such realizations, the entries of β indexed by $S_{\boldsymbol{\ell}}$ (i.e., locations where $b_{S_{\boldsymbol{\ell}}^c}$ equals $\star$) are uniform on the set of all possible subsequences that are consistent with π, of which there are $|\mathcal{B}_{\boldsymbol{\ell}}(\pi)|$ in total. Hence, Fano's inequality [9, Sec. 2.10] gives

$$I(\beta; \mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c} = b_{S_{\boldsymbol{\ell}}^c}, \mathbf{X} = \mathbf{x}) \ge \log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| \cdot \big(1 - \delta(b_{S_{\boldsymbol{\ell}}^c}, \mathbf{x})\big) - \log 2, \tag{28}$$

and averaging both sides over $(\beta_{S_{\boldsymbol{\ell}}^c}, \mathbf{X})$ gives the following generalization of (27):

$$I(\beta; \mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c}, \mathbf{X}) \ge \log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| \cdot (1-\delta) - \log 2, \tag{29}$$

where we recall that δ is the target error probability. This provides the starting point of our analysis.

Step 2: Bounding the mutual information. We upper bound the conditional mutual information in (29) as

$$\begin{aligned}
I(\beta;\mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) &= H(\mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) - H(\mathbf{Y}|\beta,\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) && (30)\\
&= H(\mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) - \sum_{i=1}^n H(Y^{(i)}|\beta,\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) && (31)\\
&= H(\mathbf{Y}|\beta_{S_{\boldsymbol{\ell}}^c},\mathbf{X}) - \sum_{i=1}^n H(Y^{(i)}|\beta,\beta_{S_{\boldsymbol{\ell}}^c},X^{(i)}) && (32)\\
&\le \sum_{i=1}^n H(Y^{(i)}|\beta_{S_{\boldsymbol{\ell}}^c},X^{(i)}) - \sum_{i=1}^n H(Y^{(i)}|\beta,\beta_{S_{\boldsymbol{\ell}}^c},X^{(i)}) && (33)\\
&= \sum_{i=1}^n I(\beta;Y^{(i)}|\beta_{S_{\boldsymbol{\ell}}^c},X^{(i)}), && (34)
\end{aligned}$$

where:

- (31) follows since we have assumed that the observations are conditionally independent;
- (32) follows since given $(\beta,\beta_{S_{\boldsymbol{\ell}}^c})$, $Y^{(i)}$ depends on $\mathbf{X}$ only through $X^{(i)}$;
- (33) follows from the sub-additivity of entropy and the fact that conditioning reduces entropy (e.g., see [9, Ch. 2]).

Substituting (34) into (29), re-arranging, and maximizing over $\boldsymbol{\ell}$ (which was arbitrary in the above analysis), we obtain Theorem 2.

B.2 Proof of Corollary 1

In the case that the entries of $\mathbf{X}$ are i.i.d. Bernoulli, each mutual information term $I(\beta;Y^{(i)}|\beta_{S_{\boldsymbol{\ell}}^c},X^{(i)})$ is identical, and (10) becomes

$$n \ge \max_{\boldsymbol{\ell}\,:\,\|\boldsymbol{\ell}\|_0\ge 2} \frac{\big(\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)|\big)(1-\delta)-\log 2}{I(\beta;Y|\beta_{S_{\boldsymbol{\ell}}^c},X)}, \tag{35}$$

where we define $(X,Y)=(X^{(i)},Y^{(i)})$ for some arbitrary fixed $i\in\{1,\dots,n\}$.

Let $X_{0,\boldsymbol{\ell}}$ (respectively, $X_{1,\boldsymbol{\ell}}$) be formed from X by taking the sub-vector of X indexed by $S_{\boldsymbol{\ell}}$ (respectively, $S_{\boldsymbol{\ell}}^c$) and re-ordering it so that the indices corresponding to class 1 appear first, then class 2, and so on. Since the entries of X are i.i.d. Bernoulli, this means that the triplet $(X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}},Y)$ follows the joint distribution described in Theorem 2. We have

$$\begin{aligned}
I(\beta;Y|\beta_{S_{\boldsymbol{\ell}}^c},X) &= H(Y|\beta_{S_{\boldsymbol{\ell}}^c},X) - H(Y|\beta,\beta_{S_{\boldsymbol{\ell}}^c},X) && (36)\\
&= H(Y|X_{1,\boldsymbol{\ell}},\beta_{S_{\boldsymbol{\ell}}^c},X) - H(Y|X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}},\beta,\beta_{S_{\boldsymbol{\ell}}^c},X) && (37)\\
&\le H(Y|X_{1,\boldsymbol{\ell}}) - H(Y|X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}},\beta,\beta_{S_{\boldsymbol{\ell}}^c},X) && (38)\\
&= H(Y|X_{1,\boldsymbol{\ell}}) - H(Y|X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}}) && (39)\\
&= I(X_{0,\boldsymbol{\ell}};Y|X_{1,\boldsymbol{\ell}}), && (40)
\end{aligned}$$

where:

- (37) follows since $X_{1,\boldsymbol{\ell}}$ is a function of $(\beta_{S_{\boldsymbol{\ell}}^c},X)$ and $(X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}})$ is a function of $(\beta,\beta_{S_{\boldsymbol{\ell}}^c},X)$;
- (38) follows since conditioning reduces entropy;
- (39) follows since Y and $(\beta,\beta_{S_{\boldsymbol{\ell}}^c},X)$ are conditionally independent given $(X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}})$. This is because the model is of the form (3), and the values $\{N_t\}_{t=1}^d$ are already determined by $(X_{0,\boldsymbol{\ell}},X_{1,\boldsymbol{\ell}})$.

Substituting (40) into (35) completes the proof.

C Applications of Theorem 2 to specific models

C.1 Noiseless setting with arbitrary testing

Here we prove the first claim given in the application to the noiseless model following Theorem 2.

In the noiseless setting, $Y^{(i)}$ is a deterministic function of $(\beta, X^{(i)})$, and hence $I(\beta; Y^{(i)} | \beta_{S_{\boldsymbol{\ell}}^c}, X^{(i)}) = H(Y^{(i)} | \beta_{S_{\boldsymbol{\ell}}^c}, X^{(i)})$. It turns out to suffice to let $\boldsymbol{\ell} = (\ell_1, \dots, \ell_d)$ be such that each ℓ_t either equals its minimum value zero or its maximum value $p\pi_t$. We let $G \subseteq [d]$ index those equaling the maximum value, and let $G^c = [d] \backslash G$ index those equaling zero. As a result, $\beta_{S_{\boldsymbol{\ell}}^c}$ in (9) is precisely equal to β_{G^c} in (25), and we are left to bound $H(Y^{(i)} | \beta_{G^c}, X^{(i)})$. For notational simplicity, we focus on an arbitrary fixed value of i and omit the superscripts $(\cdot)^{(i)}$.

Recall that $Y = (Y_1, \dots, Y_d)$ according to (2). Given G, we let G' be an arbitrary subset of G with a single element removed, and we write $Y_G = (Y_t)_{t \in G}$, and similarly for Y_{G^c} and $Y_{G'}$. With these definitions, we have

$$\begin{aligned} H(Y | \beta_{G^c}, X) &= H(Y_G, Y_{G^c} | \beta_{G^c}, X) && (41) \\ &= H(Y_{G'}, Y_{G^c} | \beta_{G^c}, X) && (42) \\ &= H(Y_{G'} | \beta_{G^c}, X) && (43) \\ &\leq \sum_{t \in G'} H(Y_t | \beta_{G^c}, X), && (44) \end{aligned}$$

where (42) follows since any single entry of Y can be uniquely determined as equaling the number of ones in X minus the other $d-1$ entries, (43) follows since Y_{G^c} is deterministic given (β_{G^c}, X), and (44) follows from the sub-additivity of entropy.

We proceed by characterizing the conditional distribution of Y_t for given values of (β_{G^c}, X). Let m_G denote the total number of ones in X among the indices where β_{G^c} equals $\star$ (i.e., the indices of items whose labels are in G). We denote these indices by S_G. Moreover, recall that β is uniform on $\mathcal{B}(\pi)$, so once β_{G^c} is known, the remaining entries are uniform on the set of possible outcomes consistent with both π and $\beta_{S_{\boldsymbol{\ell}}^c}$.

From these definitions and observations, we see that the items within S_G having label t are obtained by randomly selecting $\pi_t p$ indices uniformly at random without replacement from a total of $p_G := \sum_{t \in G} \pi_t p$ indices. Since Y_t represents the number of such locations where X equals one, we have

$$(Y_t | \beta_{G^c}, X) \sim \text{Hypergeometric}(\pi_t p, m_G, p_G). \tag{45}$$

Note that for $G = [d]$, this matches the distribution derived in Appendix A. Before proceeding, we present the following lemma regarding the entropy of an integer-valued random variable.

Lemma 2. [18] *For any integer-valued random variable U, we have*

$$H(U) \leq \frac{1}{2} \log \bigg(2\pi e \Big(\mathrm{Var}[U] + \frac{1}{12} \Big) \bigg). \tag{46}$$

Note that the right-hand side of (46) is the *differential entropy* of a Gaussian random variable with variance $\mathrm{Var}[U] + \frac{1}{12}$ [9, Ch. 8]. For continuous random variables, an analogous result holds true without the addition of $\frac{1}{12}$, i.e., the Gaussian distribution maximizes entropy for a given variance.

For $U \sim$ Hypergeometric(k, m, p), we have

$$\mathrm{Var}[U] = k \cdot \frac{m}{p} \cdot \frac{p-m}{p} \cdot \frac{p-k}{p-1} \leq \frac{k}{4}, \tag{47}$$

where we have applied $\frac{p-k}{p-1} \leq 1$ and $m(p-m) \leq \frac{p^2}{4}$. Hence, under the distribution in (45), the conditional variance of Y_t is upper bounded by $\pi_t p/4$, and Lemma 2 yields

$$\begin{aligned} H(Y_t | \beta_{G^c}, X) &\leq \frac{1}{2} \log \bigg(2\pi e \Big(\frac{\pi_t p}{4} + \frac{1}{12} \Big) \bigg) && (48) \\ &= \Big(\frac{1}{2} \log p \Big) (1 + o(1)), && (49) \end{aligned}$$

where in (49) we used the fact that π does not depend on p (and hence $\pi_t = \Theta(1)$ for all t). Substituting (49) into (44) and noting that $|G'| = |G|-1$, we obtain $H(Y|\beta_{G^c}, X) \le \big(\frac{|G|-1}{2}\log p\big)(1+o(1))$.

Putting it all together, we have shown that $I(\beta; Y^{(i)}|\beta_{G^c}, X^{(i)}) \le \big(\frac{|G|-1}{2}\log p\big)(1+o(1))$ for all $i = 1, \dots, n$. In addition, we have analogously to (21) that $\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| = p_G(H(\pi_G) + o(1))$ under our choice of $\boldsymbol{\ell}$ (depending on G). Hence, substituting into (10), maximizing over G, and changing variables from $|G|$ to r analogously to Section 3.1, we obtain (8) with $1-\delta-o(1)$ in place of $1-\eta$.

C.2 Noiseless setting with Bernoulli testing

Here we derive (11) for the noiseless model with Bernoulli testing. We follow the same arguments as those used in Section C.1 for general tests, and therefore only describe the differences. We restrict the choices of $\boldsymbol{\ell}$ as in Section C.1, indexing them by $G \subseteq [d]$ and using the definition of β_{G^c} in (25). Moreover, we write (X_G, X_{G^c}) in place of $(X_{0,\boldsymbol{\ell}}, X_{1,\boldsymbol{\ell}})$.

The mutual information term $I(X_G; Y|X_{G^c})$ simplifies to $H(Y|X_{G^c})$ in the noiseless setting, and analogously to (44), we have

$$H(Y|X_{G^c}) = \sum_{t\in G'} H(Y_t|X_{G^c}), \tag{50}$$

where G' is an arbitrary subset of G with a single element removed. Next, we observe that each Y_t for $t \in G'$ is in fact independent of X_{G^c}, and is distributed as $\mathrm{Binomial}(p\pi_t, q)$. The corresponding variance is $p\pi_t q(1-q)$, and applying Lemma 2, we conclude that the entropy is upper bounded by $\frac{1}{2}\log\big(2\pi e\big(p\pi_t q(1-q) + \frac{1}{12}\big)\big)$. Since we have assumed that pq and $p(1-q)$ both grow unbounded as $p\to\infty$, and recalling that $\pi_t = \Theta(1)$, this simplifies to $\big(\frac{1}{2}\log(pq(1-q))\big)(1+o(1))$.

Once this upper bound on the entropy of each Y_t is established, we deduce (12) using (11) and the same argument as that following (49).

C.3 Gaussian noise with large signal-to-noise ratio

Here we derive the first bound (15) for the Gaussian noise model.

We again restrict the choices of $\boldsymbol{\ell}$ as in Section C.1, indexing them by $G \subseteq [d]$ and using the definition of β_{G^c} in (25). Letting $H(\cdot)$ denote the differential entropy [9, Ch. 8] of a continuous random variable, we have

$$\begin{aligned} I(\beta; Y|\beta_{G^c}, X) &= H(Y|\beta_{G^c}, X) - H(Y|\beta, \beta_{G^c}, X) && (51)\\ &= H(Y|\beta_{G^c}, X) - \frac{d}{2}\log(2\pi e p\sigma^2) && (52)\\ &\le \sum_{t=1}^{d} H(Y_t|\beta_{G^c}, X) - d\log(2\pi e p\sigma^2), && (53)\\ &= \sum_{t\in G} H(Y_t|\beta_{G^c}, X) - (d - |G^c|)\log(2\pi e p\sigma^2), && (54) \end{aligned}$$

where

- (52) follows since the only uncertainty in Y given (β, β_{G^c}, X) is that of the d additive $\mathrm{N}(0, p\sigma^2)$ terms, each of which has differential entropy $\frac{1}{2}\log(2\pi e p\sigma^2)$ [9, Ch. 8];
- (53) follows from the sub-additivity of entropy;
- (54) follows since for $t \in G^c$, the only uncertainty in Y_t given (β_{G^c}, X) is that of the additive $\mathrm{N}(0, p\sigma^2)$ noise term.

For $t \in G$, each Y_t is of the form $N_t + Z_t$, where N_t is (conditionally) distributed as in (45), and $Z_t \sim \mathrm{N}(0, p\sigma^2)$ is independent of N_t. Using (47), we deduce that $\mathrm{Var}[Y_t|\beta_{G^c}, X] \le p\pi_t/4 + p\sigma^2$ for any realizations of (β_{G^c}, X), which in turn implies $H(Y_t|\beta_{G^c}, X) \le \frac{1}{2}\log\big(2\pi e(p\pi_t/4 + p\sigma^2)\big)$ since the Gaussian distribution maximizes the differential entropy for a given variance [9, Thm. 8.6.5].

Substituting into (54) and noting that $d - |G^c| = |G|$, we obtain

$$I(\beta; Y|\beta_G, X) \le \sum_{t \in G} \frac{1}{2} \log\big(2\pi e(p\pi_t/4 + p\sigma^2)\big) - |G| \log(2\pi e p\sigma^2) \tag{55}$$

$$= \sum_{t \in G} \frac{1}{2} \log\bigg(1 + \frac{\pi_t}{4\sigma^2}\bigg). \tag{56}$$

In addition, as we already stated in the noiseless case, it holds that $\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| = p_G(H(\pi_G) + o(1))$ under our choice of $\boldsymbol{\ell}$ (depending on G). Substituting into (10) and maximizing over G, we obtain the desired bound in (15).

C.4 Gaussian noise with constant signal-to-noise ratio

Here we derive the second bound (16) for the Gaussian model.

We choose $\boldsymbol{\ell}$ in (8) with $\ell_1 = p\pi_1$, $\ell_2 = 1$, and $\ell_t = 0$ for $t = 3, \dots, d$. Since only $p\pi_1 + 1$ entries of β remain unspecified (i.e., the corresponding entries of $\beta_{S^c_{\boldsymbol{\ell}}}$ are equal to $\star$), and those become fully specified once we assign the single remaining item with label 2 (since this means the rest must have label 1), we have

$$|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| = p\pi_1 + 1. \tag{57}$$

The main step is to bound the mutual information terms appearing in (8). We again focus on a single test indexed by i, and write (X, Y) in place of $(X^{(i)}, Y^{(i)})$. We have

$$I(\beta; Y|\beta_{S^c_{\boldsymbol{\ell}}}, X) = I(\beta; Y_1, Y_2|\beta_{S^c_{\boldsymbol{\ell}}}, X) \tag{58}$$

$$= H(Y_1, Y_2|\beta_{S^c_{\boldsymbol{\ell}}}, X) - H(Y_1, Y_2|\beta, \beta_{S^c_{\boldsymbol{\ell}}}, X) \tag{59}$$

$$= H(Y_1, Y_2|\beta_{S^c_{\boldsymbol{\ell}}}, X) - \log(2\pi e(p\sigma^2)) \tag{60}$$

$$\le H(Y_1|\beta_{S^c_{\boldsymbol{\ell}}}, X) + H(Y_2|\beta_{S^c_{\boldsymbol{\ell}}}, X) - \log(2\pi e(p\sigma^2)), \tag{61}$$

where

- (58) follows since $Y_3, \dots, Y_d$ are conditionally independent of β given $(\beta_{S^c_{\boldsymbol{\ell}}}, X)$ (specifically, they are pure Gaussian noise due to the choice of $\boldsymbol{\ell}$);
- (60) follows since Y_1 and Y_2 are also pure Gaussian noise given $(\beta, \beta_{S^c_{\boldsymbol{\ell}}}, X)$, so they each have entropy $\frac{1}{2}\log(2\pi e(p\sigma^2))$;
- (61) follows from the sub-additivity of entropy.

To bound $H(Y_1|\beta_{S^c_{\boldsymbol{\ell}}}, X)$, we recall that $Y_1 = N_1 + Z_1$, where N_1 counts the number of tested items with label 1, and $Z_1 \sim \mathrm{N}(0, p\sigma^2)$. We write this as $Y_1 = N_{\text{total}} - \xi + Z_1$, where N_{total} is the total number of unspecified items included in the test (i.e., the number of $j \in [p]$ such that $(\beta_{S^c_{\boldsymbol{\ell}}})_j = \star$ and $X_j = 1$), and $\xi \in \{0, 1\}$ indicates whether the single unspecified item with label 2 is tested.

Since the quantity N_{total} is deterministic given $(\beta_{S^c_{\boldsymbol{\ell}}}, X)$, the conditional variance of Y_1 is simply

$$\mathrm{Var}[Y_1|\beta_{S^c_{\boldsymbol{\ell}}}, X] = \mathrm{Var}[-\xi + Z_1] \tag{62}$$

$$= \mathrm{Var}[\xi] + \mathrm{Var}[Z_1] \tag{63}$$

$$\le \frac{1}{4} + p\sigma^2, \tag{64}$$

where (63) follows since ξ and Z_1 are independent, and (64) follows since a random variable on $\{0, 1\}$ has variance at most $\frac{1}{4}$, and since Z_1 is Gaussian with variance $p\sigma^2$. Finally, since the Gaussian distribution maximizes entropy for a given variance, we deduce that

$$H(Y_1|\beta_{S^c_{\boldsymbol{\ell}}}, X) \le \frac{1}{2} \log\bigg(2\pi e\Big(\frac{1}{4} + p\sigma^2\Big)\bigg). \tag{65}$$

For Y_2, we apply the same argument, noting that $Y_2 = N_{2,\text{other}} + \xi + Z_2$, where $N_{2,\text{other}}$ counts the number of indices where $(\beta_{S^c_{\boldsymbol{\ell}}})_j = 2$ and $X_j = 1$. We see that $N_{2,\text{other}}$ is deterministic given

$(\beta_{S_{\boldsymbol{\ell}}^c}, X)$, and it follows that Y_2 satisfies the same conditional variance bound as Y_1, and hence the same conditional entropy bound as (65).

Substituting (65) (and the analog for Y_2) into (61), we obtain

$$I(\beta; Y | \beta_{S_{\boldsymbol{\ell}}^c}, X) \le \log(2\pi e(1/4 + p\sigma^2)) - \log(2\pi e(p\sigma^2)) \tag{66}$$

$$= \log\Big(1 + \frac{1}{4p\sigma^2}\Big) \tag{67}$$

$$\le \frac{1}{4p\sigma^2}, \tag{68}$$

where (68) follows from the inequality $\log(1+\alpha) \le \alpha$. Finally, substituting (57) and (68) into (10) and writing $\log(p\pi_1 + 1) = (\log p)(1 + o(1))$, we obtain the desired result (16).

D Extensions to approximate recovery

Throughout the paper, we have considered the exact recovery criterion in (4), in which one insists on estimating every entry of β correctly. However, both Theorems 1 and 2 extend readily to the *approximate recovery* setting, as we describe below. We note that relaxed recovery criteria are known to considerably reduce the number of measurements in certain problems such as compressive sensing [19, 20], while having a smaller effect in other problems including group testing [16, 21].

Suppose that we only require the recovery of β up to a Hamming distance of $q_{\max} \in \{0, \dots, p\}$. Then the error probability is given by

$$P_{\mathrm{e}}(q_{\max}) = \mathbb{P}\bigg[\sum_{j=1}^{p} \mathbb{1}\{\hat{\beta}_j \ne \beta_j\} > q_{\max}\bigg]. \tag{69}$$

One should certainly expect this criterion to reduce the number of measurements for certain values of $q_{\max}$: If $d = 2$ and $q_{\max} \ge \max\{p\pi_1, p\pi_2\}$ then we can achieve $P_{\mathrm{e}}(q_{\max}) = 0$ with no tests, by simply declaring each entry of $\hat{\beta}$ to equal the most common label.

Nevertheless, the following generalization of Theorem 1 reveals that in the noiseless setting, the asymptotic reduction in the number of tests is insignificant when $q_{\max}$ is not too large.

Theorem 3. (Approximate recovery, noiseless) *Consider the noiseless pooled data problem under the approximate recovery criterion* (69)*, with a given number of labels d and label proportion vector π (not depending on the dimension p), and a given maximum Hamming distance $q_{\max}$. Then for any decoder, in order to achieve $P_{\mathrm{e}}(q_{\max}) \not\to 1$ as $p \to \infty$, it is necessary that*

$$n \ge \frac{1}{\log p}\bigg(\max_{r \in \{1,\dots,d-1\}} \frac{2\big(pH(\pi) - pH(\pi^r) - \log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j\big)}{d - r}\bigg)(1 - \eta) \tag{70}$$

for arbitrarily small $\eta > 0$.

Proof. The proof is identical to that of Theorem 1 up until (19), at which point the condition $\hat{\beta}(\mathbf{Y}(b)) = b$ should be replaced by $d_{\mathrm{H}}(\hat{\beta}(\mathbf{Y}(b)), b) \le q_{\max}$, where d_{H} denotes the Hamming distance. The number of sequences within a Hamming distance $q_{\max}$ of a given $b \in [d]^p$ is upper bounded by $\sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j$, which follows by counting the number of ways of choosing $j \le q_{\max}$ locations and assigning one of $d - 1$ new values to each.

As a result, the right-hand side of (20) needs to be multiplied by $\sum_{j=0}^{\alpha^* p} \binom{p}{j}(d-1)^j$, and following the remainder of the proof with this factor incorporated, we obtain (70). □

For any $q_{\max} = o(p)$, the term $\log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j$ is dominated by $pH(\pi) - pH(\pi^r)$, and hence the approximate recovery threshold is identical to the exact recovery threshold. Hence, a key implication of Theorem 3 is that asymptotically, recovering all labels is essentially as easy as recovering all but a vanishing fraction of the labels.

In contrast, if $q_{\max} = \alpha^* p$ for fixed $\alpha^* \in (0,1)$, the term $\log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j$ behaves as $\Theta(p)$, and Theorem 3 indicates that the approximate recovery criterion may permit improved constant factors in the required number of tests. However, the scaling laws are unchanged when α^* is sufficiently small (in particular, small enough to avoid the above-mentioned trivial cases).

Theorem 2 also extends naturally to the approximate recovery criterion, yielding the following.

Theorem 4. (Approximate recovery, noisy) *Consider the pooled data problem under a general observation model of the form* (3) *and the approximate recovery criterion* (69)*, with a given number of labels d, label proportion vector π, and maximum Hamming distance $q_{\max}$. Then for any decoder, in order to achieve $P_{\rm e}(q_{\max}) \le \delta$ for a given $\delta \in (0,1)$, it is necessary that*

$$n \ge \max_{\boldsymbol{\ell} \,:\, \|\boldsymbol{\ell}\|_0 \ge 2} \frac{\big(\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| - \log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j\big)(1-\delta) - \log 2}{\frac{1}{n}\sum_{i=1}^n I(\beta; Y^{(i)} | \beta_{S_{\boldsymbol{\ell}}^c}, X^{(i)})}. \tag{71}$$

Proof. The proof is nearly identical to that of Theorem 2, except that we replace Fano's inequality by its counterpart for approximate recovery, analogously to previous works on problems such as support recovery [20, Appendix A] and graphical model selection [22, Lemma 1] (see also [23]). Similarly to the proof of Theorem 3, the term $\log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j$ represents the number of different $\hat{\beta}$ that remain feasible given that β is fixed and an error does not occur. □

An approximate recovery analog of Corollary 1 follows naturally from Theorem 4, as do bounds of the form (12)–(15) with analogous modifications to those given in (70).

On the other hand, Theorem 4 does not recover any meaningful analog of (16). This is because the proof of (16) is based on a choice of $\boldsymbol{\ell}$ with $\log|\mathcal{B}_{\boldsymbol{\ell}}(\pi)| \le \log p$, which is dominated by $\log \sum_{j=0}^{q_{\max}} \binom{p}{j}(d-1)^j$ in (71) unless $q_{\max} = 0$. Stated differently, the proof of (16) essentially involves leaving the decoder with the difficulty of estimating one specific label, which is trivial in the approximate recovery setting.

Nevertheless, in the constant signal-to-noise ratio regime with either $q_{\max} = o(p)$ or $q_{\max} = \alpha^* p$ for sufficiently small $\alpha^* \in (0,1)$, one can still use the analog of (15) to prove an $\Omega(p)$ lower bound. While this is not as strong as the $\Omega(p \log p)$ bound proved for exact recovery, it still shows that noise increases the number of tests from sub-linear in the dimension to at least linear.