Robust lossy detection using sparse measurements: The regular case

Balakrishnan Narayanaswamy, Rohit Negi, Pradeep Khosla

Department of Electrical & Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract—Sparse measurement structures - in which each measurement only depends on a small number of the inputs - arise in models of many problems such as sensor networks, group testing and even lossless data compression. The most important question in these applications is ‘How many measurements are sufficient to reconstruct the input?’. Regular structures, where each input is measured the same number of times, require fewer measurements than when arbitrary measurements are used. In this paper we conduct a general analysis of the performance of these regular measurement structures when used for lossy reconstruction of the input in the presence of measurement noise. The main contribution of our work is the generality of the result which is applicable to lossy detection with arbitrary, even non-linear measurements and with inputs and outputs in any discrete domain. The second contribution is the quantification of the effect of noise in the measurements on the reconstruction performance, which can be used to analyze the robustness of these sparse measurement structures. We show the generality and applicability of our results by analyzing the performance of pooling designs for rare allele detection.

I. INTRODUCTION

The problem of detecting a discrete vector from measurements appears in many different guises in signal processing and information theory. The problems of group testing [1], communicating across a multiple access channel [2], detection in sensor networks [3], communicating over a noisy channel and many others can naturally be cast as problems of reconstructing a discrete vector from noisy measurements. Sparse measurements - where each measurement is a function of a small number of the variables to be detected - are studied either because of real world constraints or because they have efficient detection algorithms. For example, (i) Rachlin et. al. [4] model sensor measurements as being sparse, (ii) Montanari et. al. [5] suggest the use of sparsely-spread Code-Division Multiple Access (CDMA) for multiuser detection because they can be decoded efficiently (iii) Bruno et. al. [6] used low weight pools for group testing because they have better performance and (iv) Montanari et. al. [7] suggested the use of sparse measurements for lossless compression because of a smoothness property - changing one input changes only a few outputs.

We are interested in sparse measurement structures as a way to model the design of ‘smart’ pooling designs used when screening for rare genetic variants [8], [9]. Many disorders (e.g. sickle-cell anemia [10]) are caused by changes in specific locations in the genome and can be detected by sequencing small parts of the genome. Since the prevalence rate of these alternate alleles is low, it is possible to pool together samples from a large number of individuals and sequence the pools to reduce the number of tests required. The results from the pools are then decoded to detect carriers and affected individuals [8]. This is reminiscent of the group testing problem [1], except that the input alphabet is ternary (as opposed to binary in the group testing problem) and that the measurement function is a weighted sum (as opposed to a binary OR). Parallels have also been made to the problem of compressed sensing in [9]. However, inputs in the compressed sensing problem are continuous and the suitability of the pooling matrices for compressed sensing problems is hard to test. In addition, the peculiar noise model that arises in Next-Generation Sequencers is hard to incorporate. In this work we try to rectify these model mismatches by conducting an information theoretic study of the performance of general measurements, over any discrete alphabet with arbitrary noise models.

Based on parallels to information and communication theory we use a random measurements analog of Shannon’s random coding argument and derive bounds on the achievable rates (the number of measurements sufficient) for lossy detection with noisy measurements. Since sparse measurements are not an ideal code, bounds on achievable performance are not obvious and bounding techniques need to be modified to account for correlations between the measurements [4], [11]. Our final result can be stated in terms of a single measurement. Thus, the performance of a large measurement network, like the one in Fig. 1 (for large n), can be reduced to a calculation on the single measurement factor graph in Fig. 2.

In order to state our main result, we outline our model. We study the problem of detecting k discrete random variables (inputs, \( \bar{V} = [V_1, \ldots, V^k] \)) (each \( V^j \)) drawn independently, identically distributed (i.i.d) from \( P_{V^j} \), from \( n = \frac{k}{R} \) (where \( R \) is the ‘rate’) noisy, sparse and location-regular measurements \( \bar{Y} = [Y_1, \ldots, Y^n] \). An example network is shown in the factor graph notation [12] in Fig. 1. We study the problem in the large system limit as \( n \to \infty \) and answer the question, ‘What rates \( R \) are sufficient to detect the input to within a specified distortion using sparse location-regular noisy measurements?’.

Theorem 1.1: Lower bound on capacity of location-regular measurements: There exists a sequence of sparse, location-regular measurement networks \( \bar{C}_n \), such that the probability that a Maximum Likelihood decoder decodes to a vector that is at an average distance \( \geq D \) from the true input, \( P_{e,C_n} \to 0 \).
as $n \to \infty$, for any $R < C_{LB}(D)$, for

$$C_{LB}(D) = \inf_{P_{Z,Z'}: \frac{1}{n} \mathbb{E}_{d}(Z,Z') \geq D} \left( \frac{cH(Z^j, Z'^j) - H(Z, Z') + H(Y|Z') - H(Y|Z)}{H(Z^j, Z'^j) - H(Z^j)} \right)$$

where, referring to Fig. 2, the $Z = [Z^1, \ldots, Z^n] \in \mathcal{V}$ with entries drawn i.i.d from $P_{Z} = P_V$ - the true distribution of the inputs - for $i = 1, \ldots, c$. $P_{Z,Z'}$ is the joint distribution between $Z$ and $Z'$. The $Z$ is passed through the measurement function $\Psi$ and noise corruption, $P_{Y|\Psi(Z)}$, resulting in the noisy observed measurement, $Y$. $J$ is an integer random variable distributed uniformly between 1 and $c$ so that the $(Z^j, Z'^j)$ is the marginal of $P_{Z,Z'}$ consistent with $P_Y$ (see Section III-A). $H(\bullet)$ is the entropy function.

**Interpretation of the result**: The $Z$ in Fig. 2 plays the role of the true input to the measurement network $\mathcal{V}$ in Fig. 1. The $P_{Z,Z'}$ that optimizes the bound indicates the ‘most confusing’ input pairs. The $[cH(Z^j, Z'^j) - H(Z, Z')]$ term corresponds to the (average log) probability of ‘seeing’ this $(Z, Z')$ type, over the random choice of measurement network. The $H(Y|Z') - H(Y|Z)$ indicates how much the noisy measurement can help differentiate $Z$ and $Z'$ and thus quantifies the effect of noise. The $[H(Z^j, Z'^j) - H(Z^j)]$ corresponds to the number of confusible inputs at a particular distortion from the true input and (essentially) quantifies the effect of allowing some distortion in the reconstruction.

The rest of this paper derives this result. In Section II we describe the model in more detail. In Section III we outline the analysis leaving some technical steps for a longer publication [13]. In Section IV we show some applications of our result.

### II. MODEL DESCRIPTION

As stated in the Introduction, we study a model that generalizes models studied in sensor networks, group testing and LDPC codes - both by allowing for arbitrary measurement functions and by considering the effect of noise in the measurements. We first outline the model and then give two examples that show how it can be specialized to some particular applications, motivating the general problem we study, namely, lossy detection with sparse measurements.

#### A. Problem Statement

**Model**: Our general model can be represented using a factor graph [12] as in Fig. 1. The state of the input is modeled by $k$ discrete random variables $\mathcal{V} = [V^1, \ldots, V^k]$ drawn i.i.d from $P_V$ with values in $\mathcal{V}$. There are $n$ measurements, $u \in \{1, \ldots, n\}$, each of which is sparse i.e., a function of $c$ inputs $Z^u = [V^{l(u,1)}, \ldots, V^{l(u,c)}] \in \mathcal{V}$. Each input is measured the same number $(d)$ of times resulting in a measurement network $\mathcal{C}_n$. We put the $n$ $Z^u$s as entries of a vector of length $n$ forming $\bar{Z} \in (\mathcal{V})^n$. Each noiseless function value $X^u \in \mathcal{X}$, is a function $X^u = \Psi(V^{l(u,1)}, \ldots, V^{l(u,c)}) \in \mathcal{Y}$. Each measurement is corrupted independently according to a probability mass function $(p.m.f.) P_{Y|X}(Y = y|X = x)$ resulting in the observed measurements $\bar{Y} \in \mathcal{Y}$ (see also Section III-A).

**Decoder**: We use a decoder $q : \mathcal{Y}^n \to \mathcal{V}^k$ to obtain an estimate $\hat{\mathcal{V}}$ of the input. The object of our interest is $P_{e,\mathcal{C}_n}$, the probability that the output of an ML decoder is outside a tolerable distortion region of the true state of input. We define a tolerable distortion region around an input vector $\bar{v}_m$, as $D_{\bar{v}_m} = \{m' : \frac{1}{n}d_{GEN}(\bar{v}_m, m') < D\}$ for some distortion $D \in [0,1]$ where $d_{GEN}(\bullet, \bullet)$ is a general additive distance function, such as the Hamming distance $d_H$. Motivated by applications, we define an error event for a given measurement network $\mathcal{C}$ and true input $\bar{v}_m$ as the event $\{q(\bar{Y}) \notin D_{\bar{v}_m}\}$.

**Analysis**: We study the problem in the large system limit, $(n \to \infty)$, while the rate $R = \frac{k}{n}$ is maintained a constant. We study the average probability of error $P_{e,\mathcal{C}_n}$ at distortion $D$ as the maximum rate $R$ such that there exists a sequence of measurement networks $\mathcal{C}_n$, with $P_{e,\mathcal{C}_n} \to 0$ as $n \to \infty$ for fixed $R$. By definition, a larger measurement capacity implies that we can detect the input using fewer measurements. We are interested in lower bounds on the measurement capacity (achievable rates for detection).

### III. ANALYSIS

Motivated by the work of [4] we will use parallels to information and communication theory to analyze performance.
in our general model. Inspired by the successes reported in [6], [7] and [14] we will study location-regular measurement networks. To recall, in such networks, not only is each measurement a function of the same number \((c)\) of inputs, but each input is measured the same number \((d)\) times. This is similar to the \((c,d)\)-regular check matrix used in the LDPC code literature [15] however, we generalize this to arbitrary (even non-XOR) measurements and consider the effect of noise in the measurements. The analysis techniques of [4] relied critically on the independences made possible by not forcing each location to be measured the same number of times, and cannot be modified to our problem. We instead use the techniques developed in [11], and extend them further to analyze the regular case. This requires using some new combinatorial estimates and some large deviations results.

Outline: We now proceed to our analysis of the measurement capacity of sparse location-regular structures for robust lossy detection. We will analyze the performance of a ML decoder that reconstructs an estimate of the input using noisier measurements. Our analysis can be divided into two parts. The first part (Section III-B) is based on the insights from [11] and results in a general large deviation question for detection with sparse measurements. We then diverge from [11], which was for location-irregular structures and was focused on analyzing the effects of model mismatch and parameter uncertainty. In Section III-C we outline how (using some combinatorial arguments) we can extend the results of Section III-B to the case that is of interest to us - the location-regular model.

A. Random measurements and the method of types

Connections to information theory: Following [4], [11] we will use a random measurement argument paralleling Shannon’s random coding argument in his model of a communication system. The inputs \(\hat{X}_m\) are the message to be communicated, \((m \in \{1, \ldots, |V|^k\})\). A measurement network is generated according to a probability distribution, and the noiseless (unobserved) measurements \(\hat{X}_m\) are the transmitted codeword of length \(n\) with a rate \(R = \frac{k}{n}\). Each measurement, \(u \in \{1, \ldots, n\}\), is a function of \(c\) inputs \(\hat{Z}_m = \{V^{i(u,1)}, \ldots, V^{i(u,c)}\}\), forming a vector \(\hat{Z}_m \in \{0,1\}^n\). Noise corrupts the transmitted codeword resulting in the (observed) noisy measurements \(\hat{Y}\). In Shannon’s proof the codeword for each message is drawn with i.i.d symbols from a particular distribution which is optimized to result in the best possible rate. However, this is not a good model for our purposes since we want to constrain the mappings \(\hat{V}_m \rightarrow \hat{X}_m\) to be only those mappings that can be implemented using a sparse and location-regular measurement network \(C_m\). Enforcing these constraints leads to dependencies among the codewords \(\hat{X}_m\) for different inputs \(m\) and additionally among the different measurements \(X_m^{(n)}\) for the same input \(m\). This second dependency was circumvented in [4] by not forcing location-regularity. To analyze the performance while accounting for these dependencies we will use the method of types [16].

Random measurement ensemble: Motivated by parallels to coding theory we use the socket or configuration model [15] used in the analysis of LDPC codes to generate a random measurement network. In this model a measurement network is generated as follows. We create \(d \ast k\) input sockets and \(c\) outputs. Each input corresponds to \(d\) sockets and each measurement chooses \(c\) of these uniformly without replacement. The network is now created by assigning each input to the measurements that chose its socket. This results in (with some technicalities needed for multiple edges) a network drawn uniformly from all \((c, d)\) regular bipartite graphs - or in our terms a location-regular measurement network. We use a ML decoder that returns the most probable input - \(\hat{m} = \arg\max \mathcal{P}(\hat{Y}|\hat{V}_m) = \arg\max \mathcal{P}(\hat{Y}|\hat{Z}_m)\).

Types: The type of an input vector \(\hat{v}, \hat{\gamma} = (\gamma_1, \ldots, \gamma|^Vi|)\), is defined as the empirical p.m.f. of \(\hat{v}\). The joint type of two input vectors \(\hat{v}_m\) and \(\hat{v}_{m'}\), \(\hat{X} = (\lambda_{11}, \ldots, \lambda_{|Vi||Vi|})\) is a p.m.f. of size \(|V|^2\) with \(\lambda_{ab}\) denoting the fraction of locations \(\{i: \hat{v}_m^i = a, \hat{v}_{m'}^i = b\}\). Finally the joint type of \(\hat{Z}_m\) and \(\hat{Z}_{m'}\), \(\hat{R} = (\ldots, \kappa(e, f), \ldots)\) where \(e, f \in V\) and \(\kappa(e, f)\) is the fraction of outputs \(\{u: Z_u^m = e, Z_u^{m'} = f\}\). Similar definitions arise in many applications of the method of types [3], [7].

Example of types: We quickly illustrate these concepts with an example. Suppose \(k = 3\) is the number of inputs, and consider the case where we take \(n = 3\) measurements. Let \(V = \{0,1\}\) and \(m\) and \(m'\) be such that \(\hat{V}_m = [0 \ 0 \ 1]\) and \(\hat{V}_{m'} = [0 \ 1 \ 0]\). Measurement 1 measures locations 1 and 2, Measurement 2 measures locations 2 and 3 and Measurement 3 measures locations 1 and 1. Once the measurement network is fixed we define \(\hat{Z}_m = (000, 010)\) and \(\hat{Z}_{m'} = (010, 000)\). The type of \(\hat{Z}_m\) is \([1/3, 1/3, 1/3, 0]\) since there are no occurrences of 11 in \(\hat{Z}_m\). The joint type of \(\hat{Z}_m\) and \(\hat{Z}_{m'}\) is \(\kappa(00,01) = 1/3, \kappa(01,10) = 1/3, \kappa(10,00) = 1/3\) with all other entries = 0.

B. Initial analysis based on [11]

Only typical inputs matter: We want to calculate \(P_e\) - the average probability of error averaged over measurement networks, inputs and noise. As described in Section II-A, we define \(P_{e,c}\), the average probability of error when network \(C_n\) is the measurement network. Suppose that inputs are generated i.i.d with p.m.f. \(P_V(v = v) = \hat{\gamma}\). The average probability of error, \(P_e = E_{C}E_{p,c} = \sum_{m=1}^{n^h} P_{\hat{Y}}(\hat{v} = \hat{v}_m)E_{C}E_{p,m,c}\).

We divide the environments \(\hat{v}_m\) based on their types \(\hat{\gamma}\). Using basic combinatorial results from [16], there are \(2n^{R_H(\hat{\gamma})}\) input vectors of each type \(\hat{\gamma}\), where \(\approx\) means equality up to polynomial factors. Because of the input-permutation symmetry of the socket model (Section III-C), the error rate for an input \(m\) is only a function of its type \(\hat{\gamma}\), \(P_{e,\hat{\gamma}}\) and \(P_e \approx \sum_{\gamma} 2^{-nR_H(\gamma)}2^{-nR_H(\hat{\gamma})+DKL(\hat{\gamma}|\gamma)}P_{e,\hat{\gamma}}\) where \(D_{KL}\) is the KL-divergence. In the above sum, in terms for which \(\hat{\gamma} \neq \gamma\), the \(2^{-nR_KL(\hat{\gamma}|\gamma)}\) causes their contribution to go to zero as \(n \rightarrow \infty\), since there are only a polynomial number of types [16]. So, if \(P_{e,\hat{\gamma}}\) for typical inputs \(m\) of type \(\hat{\gamma}\) goes to zero exponentially, then \(P_e \rightarrow 0\).

Gallager bounds and types: \(P_{e,m}\) is the probability that the ML decoder outputs a \(m'\) with a \(\hat{v}_m\) that is at a distortion
permutation symmetries of the socket model as explained in Section III-A, we use modified Gallager bounding methods [17]. Define $D_{\vec{\gamma}}$ to be the set of inputs $m'$ at distortion $\le D$ from the true input $m$.

$$P_{e,m} = Pr \left( \bigcup_{m' \in D_{\vec{\gamma}}} \left\{ P \left( \vec{Y} | \vec{Z}_{m'} \right) > P \left( \vec{Y} | \vec{Z}_m \right) \right\} \right)$$

We divide error events based on the joint type $\vec{\lambda}$ of the true input $m$ and the error event input $m'$. $S^X_{\vec{\gamma}}$ is the set of environments at joint type $\vec{\lambda}$ with $v_m$, and $S^T(D)$ is the set of joint types at distortion $\le D$ from $\vec{\gamma}$. Following the intuition behind the 1961 Gallager-Fano bound [17], we introduce a $\vec{X}$-specific tilting distribution $Q^X(\vec{Y} | \vec{Z}_m)$ and threshold $2^T(\vec{X})$.

$$P_{e,m} \le \sum_{\vec{\lambda} \in S^T(D)} Pr \left( P \left( \vec{Y} | \vec{Z}_m \right) < 2^n T(\vec{X}) \right) + \sum_{\vec{\lambda} \in S^T(D)} Pr \left( \bigcup_{m' \in S^X_{\vec{\lambda}}} \left\{ P \left( \vec{Y} | \vec{Z}_{m'} \right) > 2^n T(\vec{X}) \right\} \right), (2)$$

Intuitively, the first term in (2) corresponds to the event that the noise is ‘bad’ in that the probability of the true input is not large enough and the second term bounds the probability that even though the noise was ‘good’, the (random) network was large enough and the second term bounds the probability that the noise is ‘bad’ in that the probability of the true input is not.

Choosing the tilting distribution : We focus on each of the two terms in (2) separately, starting with the second term. Using results from [16], $|S^T(D)| \le n|\vec{\lambda}|$ since there are only a polynomial number of joint types, and $|S^X_{\vec{\lambda}}| \le 2^n R(\vec{X}|H(\vec{X})-H(\vec{\lambda}))$. Applying these to the second term in (2) and then using the Markov inequality, we have

$$Term 2 \le n|\vec{\lambda}| \max_{\vec{\lambda} \in S^T(D)} 2^n R(\vec{X}|H(\vec{X})-H(\vec{\lambda})) \frac{2^{-n T(\vec{X})}}{Q^X(\vec{Y} | \vec{Z}_m)}$$

$$E_{\vec{Z}_m} E_{\vec{Y} | \vec{Z}_m} E_{\vec{Y}' | \vec{Z}_m} P \left( \vec{Y} | \vec{Z}_m \right) = E_{\vec{Z}_m} E_{\vec{Y} | \vec{Z}_m} (Pr \left( \vec{Y} | \vec{Z}_m \right))$$

From (3) a choice for the tilting distribution is $Q^X(\vec{Y} | \vec{Z}_m) = E_{\vec{Z}_m} (Pr \left( \vec{Y} | \vec{Z}_m \right))$. We need to show that $Q^X$ depends only on the joint type $\vec{X}$ of $m$ and $m'$. This is because of the permutation symmetries of the socket model as explained in more detail in the next subsection. This replacement results in

$$Term 2 \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ if } R < \min_{\vec{\lambda} \in S^T(D)} \frac{T(\vec{X})}{H(\vec{X})-H(\vec{\lambda})}.$$

Large Deviations : But what is $T(\vec{X})$ in the above expression? The larger $T(\vec{X})$ is chosen to be the larger the achievable rate, but a limit is placed on the threshold by the first term in (2). This limit is given by a large deviation theorem concerning the large deviations of the random variable $I_n^X = \frac{1}{n} \log \left( \frac{P \left( \vec{Y} | \vec{Z}_m \right)}{Q^X(\vec{Y} | \vec{Z}_m)} \right)$. We are interested in the large deviation probability, $Pr(I_n^X < T(\vec{X}))$, since for the first term in (2), we have $Term 1 \le n^{\frac{|\vec{\lambda}|}{n} \min_{\vec{\lambda} \in S^T(D)} Pr \left( I_n^X < T(\vec{X}) \right)$.

We need to characterize the distribution of $I_n^X$ based on the random measurement network ensemble. Using this characterization (Section III-C) we can apply the Gartner-Ellis theorem [18] as was done in [11]. The details of the large deviation calculations are in the longer version [13].

C. Characterizing the distributions

We start with $P_m(\vec{Z})$, the probability (over the measurement ensemble) of ‘seeing’ the vector $\vec{Z}$ when the input is in state $m$ out of the $|\vec{\lambda}|$ possible states. This depends on the distribution from which the measurement ensemble is drawn. We use the socket model, as described in Section III-A, which is equivalent to drawing a measurement network uniformly at random from the set of location-regular networks.

The first observation we make is that the socket ensemble is input permutation invariant i.e., permuting the labels of the input locations does not change any of the probability distributions. In other words, $P_m(\vec{Z}) = P_{\vec{\gamma}}(\vec{Z})$ depends only on the type $\vec{\gamma}$ of the input $m$. The ensemble is also output permutation invariant - $P_{\vec{\gamma}}(\vec{Z}) = P_{\vec{\gamma}}.\vec{\delta}$ depends on the type $\vec{\delta}$ of $\vec{Z}$. Using these two symmetries and some combinatorics,

$$P_m(\vec{Z}) = P_{\vec{\gamma}}.\vec{\delta} = \begin{cases} 2^{-ncH(\vec{\gamma})} & \text{\vec{\delta} consistent with } \vec{\gamma} \\ 0 & \text{otherwise} \end{cases}$$

(4)

Moving on to $P_{m,m'}(\vec{Z},\vec{Z}')$, the probability of ‘seeing’ $\vec{Z}$ when $m$ is the input and ‘seeing’ $\vec{Z}'$ when $m'$ is the input. Similar symmetry arguments show that under the socket model, this probability only depends on the joint type $\vec{X}$ of $m$ and $m'$ and the joint type $\vec{\kappa}$ of $\vec{Z}$ and $\vec{Z}'$ (we gave an example of this joint type in Section III-A)

$$P_{m,m'}(\vec{Z},\vec{Z}') = P_{\vec{\kappa}}.\vec{\rho} = \begin{cases} 2^{-ncH(\vec{\kappa})} & \text{\vec{\rho} consistent with } \vec{\kappa} \\ 0 & \text{otherwise} \end{cases}$$

(5)

Using these expressions, and some interesting large deviation calculations, as detailed in [13], we obtain Theorem 1.1.

IV. SOME RESULTS

Specializations of our results are useful in variants of the problems mentioned in Section I, such as group testing [6], sensor networks [3] or CDMA [5]. We focus on the design of smart pooling designs for genetic screening.

Detection of rare alleles : An allele is one of the alternate forms of a gene [10]. Some disorders, such as sickle-cell anemia, are recessive in that they are exhibited only in individuals who have two copies of the alternate allele. In other cases the chance of developing the disorder increases with the number of non-functional alleles. Since parents who are carriers have a substantially higher percentage of affected children, population-wide screening may be required [19]. As discussed in Section I, and explained in more detail in [8], [9], gene sequencing can be used to screen for individuals who have 1 copy (carriers) or 2 copies (affected) of the alternate/non-functioning allele. For rare alleles, the frequency of occurrence...
of the allele is low and most people are normal and have 0 copies of the alternate allele, suggesting that samples from different individuals can be pooled and sequenced - similar to the idea behind group testing [1], [6].

**Next-Generation Sequencing Technologies (NGST)** [20] are very well suited to this purpose as they can be used to sequence many short DNA fragments in parallel ‘lanes’ and can provide a large number of ‘reads’ per lane. Each lane corresponds to a pool, and it is of practical interest to minimize the number of pools while maintaining accuracy in the presence of sampling bias and noise. The difficulty of setting up the pools and the sampling bias within pools leads to a natural constraint - each individual sample should participate in few pools resulting in sparse pooling designs. This fits into our general model as follows. The input alphabet is \( V_i \subseteq \{0, 1, 2, 3\} \) corresponding to the number of alternate alleles the \( i \)th individual has and each measurement is a pool, with the (noiseless) output the average of the inputs. The sampling bias results in a noise model \( P_{Y|X} \) that is Poisson or Gamma distributed [8] depending on the sequencing technology used. This results in a model that does not fit very well into the conventional group testing setup since the input alphabet is ternary, the output is the weighted sum and inputs and noise are probabilistic.

**Evaluating the bounds**: We next try to answer the question of whether our achievable bounds are meaningful by comparing the prediction made by the bound of Theorem 1.1 with the performance of a practical \( l_1 \) decoder [21] across a range of pool sizes. In Fig. 3 we compare the performance of the sparse pooling design when the fraction of normal individuals is .9 and when it is .99. The alternate alleles are distributed according to the Hardy-Weinberg Law [10]. We define an error only when a 1 or 2 is decoded to a 0 (or vice versa), since these are the most serious errors. Note that our bound is general enough to include such non-Hamming definitions of error.

Two interesting conclusions can be drawn from Fig. 3. The first is that the bound and algorithm show the same trends across a range of model parameters. The second conclusion is that as the probability of alternate alleles falls, the \( l_1 \) decoder performs almost as well as (our analysis of) the ML decoder. This indicates that our bound can be used to design practical pooling strategies and evaluate the effectiveness of various design choices such as (i) pool size, (ii) use of bar-coding technologies or (iii) the number of reads per lane.

**ACKNOWLEDGMENT**

NSF award CNS-0347455 and an ARO award through Cylab, CMU supported this work. We also thank the reviewers whose comments substantially improved the paper.

**REFERENCES**


