Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 11(6): 18502–18524.
DOI:10.3934/math.2026752
Received: 19 March 2026
Revised: 28 May 2026
Accepted: 02 June 2026
Published: 23 June 2026

Research article

Finite-field pooling for community state inference: sample complexity bounds at the saturation point

Jin-Taek Seong*

Graduate School of Data Science, Chonnam National University, Gwangju 61186, Republic of Korea

* **Correspondence:** Email: jtseong@jnu.ac.kr; Tel: +82-62-530-5798.

Abstract: We studied the sample complexity of community state inference, in which a K-sparse latent state vector $\mathbf{x} \in \mathbb{F}_q^N$ over a known community partition is to be recovered from pooled observations $\mathbf{y} = \mathbf{A}\mathbf{x}$ over the finite field $\mathbb{F}_q$. The pooling matrix $\mathbf{A}$ has a constant row weight of d, with modeling pools formed as linear combinations of exactly d members. We derived necessary and sufficient conditions on the number of pooled observations M. Let $\alpha_t(d)$ denote the probability that a row of $\mathbf{A}$ misses a fixed set of size t. The lower bound, obtained from Fano's inequality, is governed by $\alpha_K(d)$; the upper bound, obtained under maximum a posteriori (MAP) decoding, is governed by $\alpha_{2K}(d)$, thus reflecting the worst-case overlap between two candidate supports of total size $2K$. Under the sparse-regime approximation $\alpha_{2K}(d) \approx e^{-2Kd/N}$, we identified the asymptotic saturation point $d^* = (N/(2K)) \ln q$ as the solution of $\alpha_{2K}(d) = 1/q$, at which the two bounds match to order $\Theta(K \log_q(N/K))$ with a multiplicative gap bounded by a constant $C(q)$ that satisfies $C(q) \to 2$ as $q \to \infty$. The analysis was restricted to the sparse signal regime $K \to \infty$, $K/N \to 0$, $\log q = o(K)$, and assumes noiseless observations; no practical decoder is proposed.

Keywords: community state inference; finite fields; compressive sensing; information-theoretic bounds; MAP decoding
Mathematics Subject Classification: 05C80, 62H30, 94A15, 94B05

1. Introduction

1.1. Background and motivation

In many networked systems, the latent attributes of individual members—rather than the topology of pairwise connections—are the primary quantities of interest. Examples include the infection or immunization status of individuals in epidemiological surveillance [1], the expression level of genes in a regulatory pathway [2], and the operational state of devices in a distributed sensor network [3]. In each scenario, the community structure is either known a priori or determined by the application, and the inferential task is to determine the latent state assigned to each community member from a limited number of aggregated observations.

We refer to this task as community state inference to distinguish it from classical community detection, in which the goal is to identify the community structure itself from observed pairwise interactions (e.g., modularity maximization, stochastic block models, spectral clustering) [4–6]. The latter relies on graph-topological features, whereas the present work assumes that the community membership is given and focuses on recovering a sparse state assignment over a finite alphabet from pooled linear observations. This nomenclature, adopted throughout the manuscript, reflects the algebraic—rather than topological—nature of our analysis.

In many applications, the state vector is structurally sparse: only a small fraction of members exhibit a non-zero or atypical state at any given time. In disease surveillance, only a small subset of individuals are infected; in fault diagnoses, only a few devices are malfunctioning. This motivates a K-sparse state model with $K \ll N$, for which inference from M aggregated observations can be analyzed using tools from compressive sensing [7, 8] and group testing [1, 9].

1.2. Algebraic pooling over finite fields

To accommodate multi-valued states beyond the binary case, we represent each member's state as an element of the finite field $\mathbb{F}_q$. Finite-field modeling offers three operational benefits over real-valued representations. First, addition and multiplication are exact, thus avoiding the numerical instability that arises from continuous relaxations [10]. Second, finite-field sparse recovery admits information-theoretic guarantees: necessary and sufficient conditions for exact recovery have been characterized for general q [10, 11]. Third, finite-field operations align with tools from the coding theory [12, 13], in particular, syndrome decoding of linear block codes, thus enabling the reuse of known algorithmic and combinatorial techniques.

We consider an observation model in which each pooled observation is the $\mathbb{F}_q$-linear combination of exactly d randomly selected member states, where d is the row weight of the pooling matrix $\mathbf{A}$. The constant row weight d is a key design parameter: a small d yields sparse pooling matrices that are easy to implement but limit the information per observation, whereas a large d increases the information per observation at the cost of densification and the decoding complexity. A central question of this paper is the determination of the smallest row weight, within the sparse pooling matrix regime $d = o(N)$, at which the per-observation information rate reaches its leading-order saturation.

1.3. Research gap

Several prior works have studied finite-field sparse recovery and group testing problems related to ours. Seong [10] and Das and Vishwanath [11] established information-theoretic limits for sparse recovery over $\mathbb{F}_q$ under dense random pooling matrices, in which each entry is i.i.d. uniform on $\mathbb{F}_q$. Atia and Saligrama [14] and Mazumdar [15] analyzed non-adaptive group testing under column-regular or row-regular designs; however, their analyses were largely restricted to the binary alphabet ($q = 2$). To our knowledge, the joint dependence of the information-theoretic lower bound and the maximum a posteriori (MAP) achievable upper bound on the constant row weight d and the field size q has not been characterized in the finite-field, constant row weight pooled observation setting. The following three questions remain open:

(Q1) For a constant row weight d, how do the information-theoretic lower bound and the MAP-achievable upper bound on M jointly depend on d and q?

(Q2) What is the smallest row weight d, within the sparse pooling matrix regime $d = o(N)$, at which the MAP-achievable upper bound $M_{UB}(d)$ enters its leading-order saturation, and is the resulting sample complexity of the same order as the information-theoretic lower bound?

(Q3) Is the resulting sample complexity of order $\Theta(K \log_q(N/K))$, the scaling that arises in finite-field sparse inference [10, 11]?

The present paper addresses these questions in the asymptotic regime $N \to \infty$, $K \to \infty$, $K/N \to 0$, and $N/K \to \infty$, with $\log q = o(K)$, under a noiseless uniform-prior model.

1.4. Problem statement

We consider N members whose states form a K-sparse vector $\mathbf{x} \in \mathbb{F}_q^N$, with support drawn uniformly at random from $\binom{[N]}{K}$ and non-zero entries drawn i.i.d. uniformly from $\mathbb{F}_q \setminus \{0\}$. The observation vector $\mathbf{y} \in \mathbb{F}_q^M$ is generated as $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{A} \in \{0, 1\}^{M \times N}$ is a random pooling matrix with i.i.d. rows, each of a constant row weight of exactly d. The binary entries of $\mathbf{A}$ are mapped into $\mathbb{F}_q$ by the canonical embedding $\{0, 1\} \hookrightarrow \mathbb{F}_q$; all arithmetic in $\mathbf{y} = \mathbf{A}\mathbf{x}$ is performed over $\mathbb{F}_q$. The question we address is as follows:

> *Given N, K, q, and a target error probability $\epsilon \in (0, 1)$, what is the smallest number of pooled observations M such that exact inference of $\mathbf{x}$ from $(\mathbf{A}, \mathbf{y})$ is achievable with probability at least $1 - \epsilon$, and how does this minimum scale with d?*

1.5. Contributions

We establish a matched pair of information-theoretic bounds on the sample complexity of finite-field community state inference under the constant row weight d, and we identify a row weight at which the two bounds match in order.

(C1) *Row weight dependent lower bound (Theorem 4.3).* We derive an information-theoretic lower bound on M via Fano's inequality, in which the leading-order per-observation information rate is $(1 - \alpha_K(d)) \log_2 q$, with $\alpha_K(d) = \binom{N-K}{d} / \binom{N}{d}$ denoting the probability that a row of the pooling matrix misses the state support. This bound characterizes the dependence of the necessary sample size on the row weight, which is not present in the bounds of [10, 11] (dense matrices) or of [14] (binary alphabet).

(C2) *MAP-achievable upper bound and bottleneck identification (Theorem 5.7).* Under MAP decoding, we derive an upper bound on the error probability via a union bound over all candidate vectors at a Hamming distance $t \in [1, 2K]$ from the true state. We identify $t = 2K$ as the combinatorial bottleneck—error events which arises from candidates whose support is disjoint from that of $\mathbf{x}$—and show that the resulting upper bound on M involves the per-observation information rate $\log_2(q/(1 + \alpha_{2K}(d)(q - 1)))$.

(C3) *Saturation point of d and order-matched bounds (Proposition 6.2, Theorem 6.4).* Within the sparse pooling matrix regime $d = o(N)$, the per-observation information rate $D_{UB}(d)$ is monotonically increasing in d and exhibits diminishing returns once $\alpha_{2K}(d) \lesssim 1/q$. By defining the saturation point d^* as the unique positive real solution of $\alpha_{2K}(d) = 1/q$ under the continuous extension $\alpha_{2K}(d) := e^{-2Kd/N}$ (Proposition 6.2), we obtain the closed form $d^* = (N/(2K)) \ln q$. At $d = d^*$, both bounds satisfy $M = \Theta(K \log_q(N/K))$, and their ratio is asymptotically bounded by a constant $C(q)$ only depending on q that tends to 2 as $q \to \infty$ (Theorem 6.4).

At the order level, The resulting sample complexity $M = \Theta(K \log_q(N/K))$ matches, , the scaling that arises in th e finite-field sparse recovery [10,11], and is here attained with a constant row weight, rather than a fully dense, pooling matrix. We emphasize that our results characterize the sample complexity order; constructive equivalence to any specific code family (e.g., syndrome decoding of a particular linear block code) is not claimed.

1.6. Scope and limitations

To delimit the scope of this paper, we note the following.
(i) Information-theoretic, not algorithmic. This work establishes sample complexity limits to recover a K-sparse vector $\mathbf{x} \in \mathbb{F}_q^N$ from M constant row weight pooled observations $\mathbf{y} = \mathbf{A}\mathbf{x}$. The contributions are a converse bound (Theorem 4.3) via Fano's inequality and an achievability bound (Theorem 5.7) via MAP decoding, together with their order-match (Theorem 6.4). We do *not* propose, analyze, or benchmark any practical decoding algorithm.
(ii) MAP as a theoretical benchmark. MAP decoding is invoked as the optimal decoder for the achievability proof; its combinatorial complexity over $\mathcal{X}_K$ makes it computationally infeasible for a moderate N. MAP is only used to establish the *existence* of a decoder that achieves the stated sample complexity, following the standard methodology in information-theoretic analyses [14,16].
(iii) Algorithmic implementation is deferred. The design and analysis of computationally feasible decoders for finite-field sparse recovery is a separate research direction, reviewed in Section 2.4. Bridging our bounds to practical algorithms (e.g., quantifying the gap between the information-theoretic limit and the threshold of belief-propagation or syndrome decoders) is a follow-up direction outlined in Section 8.
(iv) Asymptotic regime. The order-match (Theorem 6.4) is established in the regime $N \to \infty$ with $K \to \infty$, $K/N \to 0$, $N/K \to \infty$, and $\log q = o(K)$, with a constant row weight of d chosen as in Proposition 6.2. Finite-N behavior is not analyzed in the present paper; Section 7 provides deterministic evaluations of the analytical bounds at $N = 1000$ solely for illustration.

1.7. Paper organization

The remainder of this paper is organized as follows. Section 2 reviews the related literature in finite-field sparse recovery, group testing, and linear block coding, and positions the present work via a comparison table (Table 1). Section 3 formalizes the observation model. Section 4 derives the information-theoretic lower bound (Theorem 4.3). Section 5 derives the MAP-achievable upper bound (Theorem 5.7). Section 6 characterizes the saturation point of the constant row weight (Proposition 6.2) and establishes the order-match of the two bounds (Theorem 6.4). Section 7 reports deterministic numerical evaluations of the bounds. Section 8 concludes the paper.

2. Related work

This section reviews three lines of research that bear on the present work: classical community detection (Section 2.1), compressive sensing and group testing in graph-structured inference (Section 2.2), and finite-field models in the coding theory (Section 2.3). Section 2.4 reviews practical decoders for context, and Section 2.5 positions the present work via a comparison table.

2.1. Classical community detection: A distinction in scope

Classical community detection seeks to recover the *partition* of nodes into communities from observed pairwise interactions. Three families of methods are widely used. Modularity maximization [4, 17] optimizes a measure of intra-cluster density relative to a random baseline; the Louvain and Leiden algorithms [18] scale to large networks but exhibit resolution limits. Spectral methods [19,20] embed nodes via the eigen-decomposition of the graph Laplacian and apply geometric clustering, with consistency results under the stochastic block model (SBM) [6]. Model-based inference [5, 21, 22] treats the partition as a latent variable in a probabilistic generative model and applies either likelihood-based or a Bayesian estimation.

These methods address a different problem from the one considered in this paper. They infer the community structure itself from edges or interaction patterns, whereas we assume the community membership is given and infer the sparse state vector from algebraic pooled observations. The graph-topological assumptions which underlies an SBM-based inference (e.g., intra- vs. inter-block edge probabilities) do not appear in our model.

2.2. Compressive sensing and group testing

Sparse signal recovery from underdetermined linear measurements has been extensively studied over the real field [7, 8], with the canonical sample complexity $M = O(K \log(N/K))$ for matrices that satisfy the restricted isometry property (RIP). Lai and McKenzie [23] adapted this framework to the graph-structured cluster recovery by exploiting the Laplacian, thus achieving $O(N \log N)$ run-time complexity under SBM-type generative models. Wang et al. [24] considered edge-sparsity in social internet of thing (IoT) networks. These works operate over the real field and exploit graph topology, whereas our analysis operates over $\mathbb{F}_q$ and treats the community structure as exogenous.

Group testing, originating in pooled medical diagnostics [9], seeks to identify a small number of "defective" items from pooled binary tests. The non-adaptive setting has been extensively studied [1]. Atia and Saligrama [14] derived information-theoretic limits for noisy non-adaptive group testing under row-regular and column-regular pooling designs over the binary alphabet. Mazumdar [15] analyzed a closely related problem under graph-constrained pools. Community-aware group testing [2, 25] exploits known community structures to reduce the number of tests. The principal difference of the present work is the extension to general q-ary alphabets, together with an analytical expression for the saturation point d^* (Proposition 6.2) as a function of N, K, and q.

2.3. Finite-field sparse recovery and coding theory

Sparse recovery over $\mathbb{F}_q$ has been studied in the coding theory and information theory literature. Seong [10] derived necessary and sufficient conditions for exact recovery of a K-sparse vector in $\mathbb{F}_q^N$

under dense i.i.d. random matrices, where each entry of $\mathbf{A}$ is uniform on $\mathbb{F}_q$. Das and Vishwanath [11] established that ℓ_0 minimization is solvable in polynomial time over finite alphabets, thereby bypassing the need for convex relaxations. These works characterize the $K \log_q(N/K)$ scaling but do not analyze the impact of the constant row weight d.

A structural analogy with the coding theory is available: a K-sparse vector $\mathbf{x} \in \mathbb{F}_q^N$ can be viewed as a K-error pattern on a length-N codeword, and the pooled observation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is the syndrome under a parity-check matrix $\mathbf{A}$ [12, 13]. Classical existence bounds on linear q-ary codes correcting K errors—most directly the q-ary Gilbert–Varshamov bound [26]—imply that $\Theta(K \log_q(N/K))$ parity checks suffice in the regime $K = o(N)$, the same order recovered in Theorem 6.4. Therefor, the present work is therefore a sample complexity counterpart of this classical bound, established under the additional structural constraint that the parity-check matrix is a constant row weight of d rather than fully dense.

2.4. Practical decoders in finite-field sparse recovery

Although the present paper does not propose a practical decoder (cf. Section 1.6), several algorithmic families address closely related decoding problems. We summarize the main approaches and their computational complexity to place our information-theoretic results in algorithmic context.

(a) Gaussian elimination. For an instance with known support, recovery reduces to solving a linear system over $\mathbb{F}_q$. Dense Gaussian elimination has complexity of $O(M^3)$; for the row-sparse pooling matrices considered here, sparse variants achieve $O(M^2 d)$ on average [27]. Support recovery without prior knowledge requires combinatorial search, thus leading to the harder problem addressed by the algorithms below.

(b) Non-binary belief propagation (NB-BP). The system $\mathbf{y} = \mathbf{A}\mathbf{x}$ admits a factor-graph representation in which each row of the pooling matrix $\mathbf{A}$ becomes a check node of degree d. Non-binary belief propagation over $\mathbb{F}_q$, developed for non-binary LDPC codes by Davey and MacKay [28] and refined in the FFT-based variant of Declercq and Fossorier [29], has per-iteration complexity of $O(Md\, q \log q)$. Empirical thresholds typically lie within a constant factor of the information-theoretic limit, but an exact gap characterization remains open for constant row weight ensembles.

(c) Syndrome decoding and information-set decoding (ISD). Reinterpreting $\mathbf{A}$ as the parity-check matrix of a linear code over $\mathbb{F}_q$ casts sparse recovery as syndrome decoding [30, 31]. For bounded-distance decoding of algebraic codes (BCH/Reed–Solomon), the Berlekamp–Massey algorithm has complexity of $O(M^2)$ but only applies to structured ensembles. For general (random) parity-check matrices the problem is NP-hard [32]; the best known sub-exponential algorithms are ISD variants whose complexity is 2^{cK} with c depending on the code rate [33].

(d) ℓ_0-minimization over finite fields. Das and Vishwanath [11] showed that exact ℓ_0-minimization over $\mathbb{F}_q$ is solvable in polynomial time when the pooling matrix satisfies a finite-field analogue of the RIP, provided the sparsity level is below an explicit threshold. Whether the constant row weight ensembles studied here satisfy this RIP-type condition is open.

(e) Approximate message passing (AMP) variants. AMP-style algorithms have been extended to finite-alphabet signals [34, 35] and to group testing matrices [36]. Their state-evolution analysis predicts thresholds that, in several special cases, match the information-theoretic bounds derived in Section 5.

Positioning. The bounds established in this paper provide a decoder-independent benchmark against which any of the above algorithms can be compared. Under the asymptotic regime of Theorem 6.4 ($N \to \infty$, $K \to \infty$, $K/N \to 0$, $N/K \to \infty$, $\log q = o(K)$), the converse (Theorem 4.3) implies that no decoder—regardless of computational cost—can recover $\mathbf{x}$ with $M = o(K \log_q(N/K))$ pooled observations under the constant row weight ensembles considered here.

2.5. *Positioning of the present work*

Table 1 summarizes the differences between the present work and the four most closely related prior contributions. Three features distinguish the present analysis: (i) the explicit treatment of the constant row weight d as an analysis parameter, with both bounds derived as functions of d; (ii) the analytical characterization of the saturation point d^* in the q-ary setting; and (iii) the order-match of the lower and upper bounds at $d = d^*$.

Table 1. Comparison of the present work with closely related prior contributions on finite-field sparse recovery and group testing.

	[10]	[11]	[14, 15]	This work
Alphabet	$\mathbb{F}_q$	$\mathbb{F}_q$	$\mathbb{F}_2$	$\mathbb{F}_q$
Pooling matrix density	Dense	Dense	Sparse (row-reg.)	Sparse (constant row weight d)
Row weight as a parameter	No	No	Implicit	Explicit
d^* characterization	—	—	Limited ($q = 2$)	Analytical (saturation)
Order-matched LB/UB	Yes (in N, K, q)	Yes (in N, K, q)	Partial	Yes (in N, K, q, d)
Decoder benchmark	ML / ℓ_0	ℓ_0	ML	MAP
Bottleneck identification	—	—	Worst-case set	$t = 2K$ (disjoint support)

Two clarifications are in order regarding the relationship to prior work. First, the dense-matrix results of [10, 11] correspond to a different ensemble (i.i.d. uniform entries on $\mathbb{F}_q$) than the constant row weight binary ensemble studied here; they are not obtained as a limit of our analysis, since our derivations rely on the sparse-regime assumption $d = o(N)$ and the exponential approximation $\alpha_t(d) = e^{-td/N}(1 + o(1))$, both of which fail as $d \to N$ (cf. Remark 6.3). Our results provide the same order, $\Theta(K \log_q(N/K))$, under a strictly sparser pooling matrix.

Second, the binary group testing analyses of [14, 15] correspond to the special case $q = 2$ but under different observation models (noisy tests, OR-based or constant row weight Boolean designs) than ours; setting $q = 2$ in Proposition 6.2 yields $d^* = (N/(2K)) \ln 2$, which is of the same order as the row weights used in those works but is not claimed to be numerically identical to any specific value stated therein.

3. System model and problem formulation

In this section, we formalize the community state inference problem over finite fields. We adopt a definition-assumption style that is consistent with standard mathematical exposition.

3.1. Definitions of the algebraic model

Definition 3.1 (Community state vector)**.** Let $\mathbb{F}_q$ denote the finite field of order q, where q is a prime power. Consider a community of N members indexed by $i \in \{1, 2, \ldots, N\}$. The community state vector is as follows:

$$\mathbf{x} = (x_1, x_2, \ldots, x_N) \in \mathbb{F}_q^N, \tag{3.1}$$

where each entry $x_i \in \mathbb{F}_q$ represents the latent state of member i.

Definition 3.2 (K-Sparse state set)**.** For a fixed integer K with $1 \leq K \leq N$, the K-sparse state set over $\mathbb{F}_q^N$ is as follows:

$$\mathcal{X}_K = \left\{ \mathbf{x} \in \mathbb{F}_q^N : \|\mathbf{x}\|_0 \leq K \right\}, \tag{3.2}$$

where $\|\mathbf{x}\|_0$ denotes the Hamming weight of $\mathbf{x}$.

Definition 3.3 (Pooling matrix)**.** A pooling matrix is a binary matrix

$$\mathbf{A} = [a_{m,i}] \in \{0, 1\}^{M \times N}, \tag{3.3}$$

where $a_{m,i} = 1$ if member i is included in pool m, and $a_{m,i} = 0$ otherwise. The binary entries $\{0, 1\}$ are identified with the corresponding elements of the prime subfield of $\mathbb{F}_q$ whenever they appear in arithmetic expressions over $\mathbb{F}_q$.

Definition 3.4 (Constant row weight)**.** A pooling matrix $\mathbf{A} \in \{0, 1\}^{M \times N}$ has a constant row weight of d if every row contains exactly d ones:

$$\sum_{i=1}^{N} a_{m,i} = d, \qquad m = 1, \ldots, M, \tag{3.4}$$

where the sum is taken over the integers.

Definition 3.5 (Observation model)**.** Given $\mathbf{A} \in \{0, 1\}^{M \times N}$ and $\mathbf{x} \in \mathbb{F}_q^N$, the observation vector $\mathbf{y} \in \mathbb{F}_q^M$ is given entrywise by

$$y_m = \sum_{i=1}^{N} a_{m,i}\, x_i \in \mathbb{F}_q, \qquad m = 1, \ldots, M, \tag{3.5}$$

where the multiplication $a_{m,i}\, x_i$ and the summation are performed in $\mathbb{F}_q$ under the canonical embedding $\{0, 1\} \hookrightarrow \mathbb{F}_q$ of Definition 3.3. Equivalently, $\mathbf{y} = \mathbf{A}\mathbf{x}$ over $\mathbb{F}_q$.

3.2. Statistical and structural assumptions

Assumption 3.6 (Sparsity of the state vector)**.** The state vector $\mathbf{x}$ belongs to $\mathcal{X}_K$.

Assumption 3.7 (Uniform prior over exactly-K-sparse vectors)**.** $\mathbf{x}$ is uniformly distributed over the subset $\{\mathbf{x} \in \mathcal{X}_K : \|\mathbf{x}\|_0 = K\}$ of $\mathcal{X}_K$; equivalently, the support $\mathrm{supp}(\mathbf{x})$ is uniformly chosen from $\binom{[N]}{K}$, and the nonzero entries are uniformly drawn i.i.d. from $\mathbb{F}_q \setminus \{0\}$.

Assumption 3.8 (Constant row weight pooling)**.** $\mathbf{A}$ has a constant row weight of d, and each row $\mathbf{a}_m$ is drawn independently, uniformly at random from $\binom{[N]}{d}$.

Assumption 3.9 (Noiseless observations)**.** The observation equation in Definition 3.5 holds deterministically holds given **A** and **x**; no additive perturbation is introduced.

Remark 3.10 (On the role of finite fields). The use of $\mathbb{F}_q$ provides exact linearity and well-defined inverses, which support algebraic decoding without the numerical instability of real-valued compressive sensing.

Example 3.11 (Illustration of the pooling operation)**.** Consider $N = 6$, $M = 3$, $d = 2$, $q = 4$, with $\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$, $\mathbf{x} = [0, 0, 0, 0, 2, 0]^\top \in \mathbb{F}_4^6$. Computing in $\mathbb{F}_4$, $y_1 = 0$, $y_2 = 2$, $y_3 = 0$, so $\mathbf{y} = (0, 2, 0)^\top$.

3.3. Inference problem and MAP decoding

Definition 3.12 (Inference problem)**.** Given $(\mathbf{A}, \mathbf{y})$, produce an estimate $\hat{\mathbf{x}} \in \mathcal{X}_K$. The error probability is as follows:

$$P_e := \Pr(\hat{\mathbf{x}} \neq \mathbf{x}), \tag{3.6}$$

under the joint law of $(\mathbf{x}, \mathbf{A})$ induced by Assumptions 3.7 and 3.8.

Definition 3.13 (MAP decoding)**.** The MAP decoder is

$$\hat{\mathbf{x}}_{\mathrm{MAP}} \in \arg\max_{\mathbf{z} \in \mathcal{X}_K} \Pr(\mathbf{x} = \mathbf{z} \mid \mathbf{y}, \mathbf{A}),$$

with ties broken arbitrarily. Under Assumption 3.7 and the noiseless model (Assumption 3.9), the posterior is uniform over the set $C(\mathbf{y}, \mathbf{A}) := \{\mathbf{z} \in \mathcal{X}_K : \mathbf{A}\mathbf{z} = \mathbf{y}\}$, so MAP decoding reduces to returning an arbitrary element of $C(\mathbf{y}, \mathbf{A})$. Consequently, the decoding-error event $\{\hat{\mathbf{x}}_{\mathrm{MAP}} \neq \mathbf{x}\}$ is contained in the event $\{\exists\, \mathbf{z} \in \mathcal{X}_K \setminus \{\mathbf{x}\} : \mathbf{A}\mathbf{z} = \mathbf{y}\}$.

Remark 3.14 (MAP as an information-theoretic benchmark). The MAP decoding attains the minimum error probability among all decoders of the form $\hat{\mathbf{x}} : (\mathbf{A}, \mathbf{y}) \mapsto \mathcal{X}_K$. We use MAP as the benchmark for the achievability bound in Section 5; computationally feasible decoders are reviewed in Section 2.4.

4. Information-theoretic lower bound

4.1. Auxiliary lemmas

Lemma 4.1 (Conditional entropy bound for a single observation)**.** *Under Assumptions 3.6–3.8, let $y_m \in \mathbb{F}_q$ denote the m-th entry of the observation vector and $\mathbf{a}_m \in \{0, 1\}^N$ as its associated row of* **A**. *Define the following:*

$$\alpha_K(d) := \frac{\binom{N-K}{d}}{\binom{N}{d}}. \tag{4.1}$$

Then,

$$H(y_m \mid \mathbf{a}_m) \;\leq\; H_2(\alpha_K(d)) + (1 - \alpha_K(d)) \log_2 q, \tag{4.2}$$

where $H_2(\cdot)$ is the binary entropy in bits, and $\log_2$ denotes the base-2 logarithm.

Proof. Condition on $\mathbf{a}_m$ with $|\mathcal{S}_m| = d$, where $\mathcal{S}_m := \{i : a_{m,i} = 1\}$. Let Z_m denote the number of nonzero entries of $\mathbf{x}$ which fall in $\mathcal{S}_m$. Under Assumption 3.7, supp($\mathbf{x}$) is uniform over $\binom{[N]}{K}$, so $|\mathrm{supp}(\mathbf{x}) \cap \mathcal{S}_m|$ follows a hypergeometric distribution and

$$\Pr(Z_m = 0 \mid \mathbf{a}_m) = \frac{\binom{N-K}{d}}{\binom{N}{d}} = \alpha_K(d).$$

Define $E_m := \mathbb{I}(Z_m \geq 1)$. By the chain rule,

$$H(y_m \mid \mathbf{a}_m) \leq H(y_m, E_m \mid \mathbf{a}_m) = H(E_m \mid \mathbf{a}_m) + H(y_m \mid E_m, \mathbf{a}_m).$$

The first term equals $H_2(\alpha_K(d))$. For the second, the following hold: if $E_m = 0$, then $y_m = 0$ deterministically, so $H(y_m \mid E_m = 0, \mathbf{a}_m) = 0$; if $E_m = 1$, then $y_m \in \mathbb{F}_q$ and $H(y_m \mid E_m = 1, \mathbf{a}_m) \leq \log_2 q$. Combining,

$$H(y_m \mid E_m, \mathbf{a}_m) \leq (1 - \alpha_K(d)) \log_2 q,$$

which yields (4.2). □

Lemma 4.2 (Logarithmic Cardinality of the Exactly-K-Sparse Stratum)**.** *For* $1 \leq K \leq N - 1$, *let* $\mathcal{X}_K^{=K} := \{\mathbf{x} \in \mathbb{F}_q^N : \|\mathbf{x}\|_0 = K\}$. *Then,*

$$\log_2 \left|\mathcal{X}_K^{=K}\right| \geq N H_2(K/N) + K \log_2(q - 1) - \tfrac{1}{2} \log_2(8K(1 - K/N)). \tag{4.3}$$

Proof. The exactly-K-sparse stratum satisfies $|\mathcal{X}_K^{=K}| = \binom{N}{K}(q - 1)^K$. The binomial bound $\binom{N}{K} \geq 2^{NH_2(K/N)}/\sqrt{8K(1 - K/N)}$ (valid for $1 \leq K \leq N - 1$) [37] yields the claim after taking $\log_2$. □

4.2. Main result: lower bound on the number of observations

Theorem 4.3 (Information-theoretic lower bound)**.** *Consider the inference problem in Definition 3.12 under Assumptions 3.6–3.9, with* $1 \leq K \leq N - 1$. *For any* $\epsilon \in (0, 1)$, *any inference scheme that achieves* $P_e \leq \epsilon$ *must satisfy the following:*

$$M \geq \frac{(1 - \epsilon)\Big[NH_2(K/N) + K \log_2(q - 1) - \tfrac{1}{2} \log_2(8K(1 - K/N))\Big] - 1}{H_2(\alpha_K(d)) + (1 - \alpha_K(d)) \log_2 q}. \tag{4.4}$$

Proof. By Assumption 3.9, $H(\mathbf{y} \mid \mathbf{x}, \mathbf{A}) = 0$, so $I(\mathbf{x}; \mathbf{y} \mid \mathbf{A}) = H(\mathbf{y} \mid \mathbf{A})$. By the chain rule and the fact that removing conditioning does not decrease entropy,

$$H(\mathbf{y} \mid \mathbf{A}) = \sum_{m=1}^{M} H(y_m \mid y_1, \ldots, y_{m-1}, \mathbf{A}) \leq \sum_{m=1}^{M} H(y_m \mid \mathbf{a}_m).$$

Applying Lemma 4.1 to each summand,

$$I(\mathbf{x}; \mathbf{y} \mid \mathbf{A}) \leq M[H_2(\alpha_K(d)) + (1 - \alpha_K(d)) \log_2 q]. \tag{4.5}$$

By Fano's inequality [16],

$$H(\mathbf{x} \mid \mathbf{y}, \mathbf{A}) \leq H_2(P_e) + P_e \log_2\Big(\left|\mathcal{X}_K^{=K}\right| - 1\Big) \leq 1 + P_e \log_2 \left|\mathcal{X}_K^{=K}\right|.$$

Since $\mathbf{x}$ is independent of $\mathbf{A}$ and uniform over $\mathcal{X}_K^{=K}$ (Assumption 3.7), $H(\mathbf{x} \mid \mathbf{A}) = \log_2 \left|\mathcal{X}_K^{=K}\right|$. Therefore,

$$I(\mathbf{x}; \mathbf{y} \mid \mathbf{A}) = H(\mathbf{x} \mid \mathbf{A}) - H(\mathbf{x} \mid \mathbf{y}, \mathbf{A}) \;\geq\; (1 - P_e) \log_2 \left|\mathcal{X}_K^{=K}\right| - 1. \tag{4.6}$$

From (4.5) and (4.6),

$$M[H_2(\alpha_K(d)) + (1 - \alpha_K(d)) \log_2 q] \geq (1 - P_e) \log_2 \left|\mathcal{X}_K^{=K}\right| - 1.$$

Using $P_e \leq \epsilon$ and Lemma 4.2, we obtain (4.4). □

Corollary 4.4 (Asymptotic form)**.** *Consider the regime $N \to \infty$ with $K/N \to 0$, $\epsilon \to 0$, and row weight $d = d(N)$ chosen so that $1 - \alpha_K(d) = \Theta(1)$. Then, the additive constant 1 in the numerator, the $\frac{1}{2}\log_2(8K(1 - K/N))$ correction, and the term $H_2(\alpha_K(d)) \in [0, 1]$ in the denominator are sub-leading relative to the dominant terms $NH_2(K/N) + K \log_2(q - 1) = \Theta(K \log(N/K))$ in the numerator and $(1 - \alpha_K(d)) \log_2 q$ in the denominator; thus,*

$$M \;\geq\; \frac{NH_2(K/N) + K \log_2(q - 1)}{(1 - \alpha_K(d)) \log_2 q} (1 - o(1)) \;-\; o(1). \tag{4.7}$$

When q is held fixed, the ratio $H_2(\alpha_K(d))/[(1-\alpha_K(d)) \log_2 q]$ only contributes a bounded multiplicative constant; when $q \to \infty$, it vanishes.

Remark 4.5 (Interpretation). The numerator of (4.7) is the logarithmic size of the K-sparse state stratum, decomposed into a spatial term $NH_2(K/N)$ (support selection) and an algebraic term $K \log_2(q-1)$ (nonzero-value selection). The denominator $(1 - \alpha_K(d)) \log_2 q$ is the per-observation information capacity: $1 - \alpha_K(d)$ is the probability that a pool intersects the support, and $\log_2 q$ is the maximum entropy of a single $\mathbb{F}_q$-valued entry of the observation vector. The bound is decoder-independent.

5. Upper bound via MAP decoding

5.1. Error event and auxiliary lemmas

Definition 5.1 (Error vector)**.** For a true state vector $\mathbf{x} \in \mathcal{X}_K$ and a candidate $\mathbf{z} \in \mathcal{X}_K$ with $\mathbf{z} \neq \mathbf{x}$, the *error vector* is

$$\mathbf{u} := \mathbf{x} - \mathbf{z} \in \mathbb{F}_q^N,$$

with a Hamming weight of $t := \|\mathbf{u}\|_0 \in [1, 2K]$ (the upper bound follows from $\|\mathbf{x}\|_0, \|\mathbf{z}\|_0 \leq K$ and the triangle inequality for $\|\cdot\|_0$).

Assumption 5.2 (Conditional uniformity of pooled inner products)**.** For any fixed nonzero error vector $\mathbf{u} \in \mathbb{F}_q^N$ of a Hamming weight of $t \geq 1$, conditioned on the event $\{\mathcal{S}_m \cap \mathrm{supp}(\mathbf{u}) \neq \emptyset\}$, the inner product $\mathbf{a}_m \cdot \mathbf{u} = \sum_{i \in \mathcal{S}_m \cap \mathrm{supp}(\mathbf{u})} u_i$ is modeled as uniformly distributed on $\mathbb{F}_q$.

Remark 5.3 (Scope of Assumption 5.2). Assumption 5.2 is a standard modeling step in finite-field group testing and sparse-recovery analyses [1, 14]. Two clarifications are in order. (i) When $|\mathcal{S}_m \cap \mathrm{supp}(\mathbf{u})| = 1$, the inner product equals a single nonzero entry of $\mathbf{u}$ and cannot equal zero. The exact conditional probability $\Pr(\mathbf{a}_m \cdot \mathbf{u} = 0 \mid \text{overlap} = 1)$ is 0, whereas Assumption 5.2 replaces it by $1/q > 0$. Since the bound on P_e derived below is monotonically increasing in $P_0(t)$, this substitution enlarges $P_0(t)$ and preserves the validity of the resulting upper bound on P_e. (ii) For overlap ≥ 2, the exact conditional distribution is a convolution of independent uniforms on $\mathbb{F}_q \setminus \{0\}$; this distribution approaches uniform on $\mathbb{F}_q$ as the overlap or q grows, but is not literally uniform at finite parameters. Hence, Assumption 5.2 is an approximation whose deviation from the exact distribution only affects constants and not the leading-order sample complexity in Theorem 5.7.

Lemma 5.4 (Single-observation nullspace probability)**.** *Let* $\mathbf{u} \in \mathbb{F}_q^N$ *be a fixed error vector of Hamming weight* $t \geq 1$*, and let* $\mathbf{a}_m$ *be a row of* $\mathbf{A}$ *generated according to Assumption 3.8. Define the following:*

$$\alpha_t(d) := \frac{\binom{N-t}{d}}{\binom{N}{d}}. \tag{5.1}$$

Then, under Assumption 5.2,

$$P_0(t) := \Pr(\mathbf{a}_m \cdot \mathbf{u} = 0) = \alpha_t(d) + (1 - \alpha_t(d))\,\frac{1}{q} = \frac{1 + \alpha_t(d)(q-1)}{q}. \tag{5.2}$$

Proof. Decompose by the overlap event:

$$\Pr(\mathbf{a}_m \cdot \mathbf{u} = 0) = \alpha_t(d) \cdot 1 + (1 - \alpha_t(d)) \cdot \Pr\!\left(\mathbf{a}_m \cdot \mathbf{u} = 0 \,\middle|\, \mathcal{S}_m \cap \operatorname{supp}(\mathbf{u}) \neq \emptyset\right).$$

The probability $\alpha_t(d)$ follows from the hypergeometric construction in Assumption 3.8; the conditional probability equals $1/q$ by Assumption 5.2. Substituting yields (5.2). □

Lemma 5.5 (Joint nullspace probability over M rows)**.** *Let* $\mathbf{u} \in \mathbb{F}_q^N$ *be a fixed nonzero error vector of weight* t*, and let* $\mathbf{A}$ *be generated under Assumption 3.8. Then, under Assumption 5.2,*

$$\Pr(\mathbf{A}\mathbf{u} = \mathbf{0}) = \prod_{m=1}^{M} \Pr(\mathbf{a}_m \cdot \mathbf{u} = 0) = [P_0(t)]^M. \tag{5.3}$$

Proof. The rows are i.i.d. by Assumption 3.8; therefore, the events $\{\mathbf{a}_m \cdot \mathbf{u} = 0\}$ are mutually independent. Each factor equals $P_0(t)$ by Lemma 5.4. □

Lemma 5.6 (Counting of candidates at error weight t)**.** *For* $t \in [1, 2K]$*, let* $\mathcal{Z}_t(\mathbf{x})$ *denote the set of* $\mathbf{z} \in \mathcal{X}_K$ *with* $\|\mathbf{x} - \mathbf{z}\|_0 = t$*. Then,*

$$|\mathcal{Z}_t(\mathbf{x})| \;\leq\; \binom{N}{t}(q-1)^t \;\leq\; \left(\frac{eN(q-1)}{t}\right)^t. \tag{5.4}$$

In particular, at $t = 2K$,

$$|\mathcal{Z}_{2K}(\mathbf{x})| \;\leq\; \left(\frac{eN(q-1)}{2K}\right)^{2K}, \qquad \log_2 |\mathcal{Z}_{2K}(\mathbf{x})| \;\leq\; 2K \log_2\!\left(\frac{eN(q-1)}{2K}\right). \tag{5.5}$$

Proof. The error vector $\mathbf{u} = \mathbf{x} - \mathbf{z} \in \mathbb{F}_q^N$ has weight t. The number of weight-t vectors in $\mathbb{F}_q^N$ equals $\binom{N}{t}(q-1)^t$, which upper-bounds $|\mathcal{Z}_t(\mathbf{x})|$ (the constraint $\mathbf{z} \in \mathcal{X}_K$ can only reduce the count). The second inequality uses $\binom{N}{t} \leq (eN/t)^t$. □

5.2. Main result: upper bound

Theorem 5.7 (MAP-Based upper bound)**.** *For any $\epsilon \in (0,1)$, MAP decoding (Definition 3.13) achieves $P_e \leq \epsilon$ provided that*

$$M \;\geq\; \max_{t\in[1,2K]} \frac{t\log_2\!\left(\dfrac{eN(q-1)}{t}\right)+\log_2(2K/\epsilon)}{\log_2\!\left(\dfrac{q}{1+\alpha_t(d)(q-1)}\right)}. \tag{5.6}$$

In the sparse regime $N \gg K$ with $d = o(N)$, the maximum is attained at $t = 2K$ at leading order, thus yielding the following specialization:

$$M \;\geq\; \frac{2K\log_2\!\left(\dfrac{eN(q-1)}{2K}\right)+\log_2(2K/\epsilon)}{\log_2\!\left(\dfrac{q}{1+\alpha_{2K}(d)(q-1)}\right)}. \tag{5.7}$$

Remark 5.8 (MAP as a theoretical benchmark). Theorem 5.7 only invokes MAP to establish the *existence* of a decoder that achieves the stated sample complexity. The combinatorial complexity of MAP over $\mathcal{X}_K$ scales with $|\mathcal{X}_K| = \sum_{k=0}^{K}\binom{N}{k}(q-1)^k$, which is super-polynomial in N for $K = \Theta(N^c)$ with $c \in (0,1)$; there, the result is an achievability statement in the standard information-theoretic sense [14, 38]. Computationally tractable decoders are surveyed in Section 2.4.

Proof of Theorem 5.7. By Definition 3.13, the MAP decoding-error event satisfies the following:

$$\{\hat{\mathbf{x}}_{\mathrm{MAP}} \neq \mathbf{x}\} \;\subseteq\; \bigcup_{\mathbf{z}\in\mathcal{X}_K\setminus\{\mathbf{x}\}} \{\mathbf{A}(\mathbf{x}-\mathbf{z}) = \mathbf{0}\} = \bigcup_{t=1}^{2K}\;\bigcup_{\mathbf{z}\in\mathcal{Z}_t(\mathbf{x})} \{\mathbf{A}\mathbf{u} = \mathbf{0}\},$$

where $\mathbf{u} = \mathbf{x}-\mathbf{z}$ has a weight of $t \in [1, 2K]$ by Definition 5.1.

By Lemmas 5.5 and 5.6,

$$P_e \;\leq\; \sum_{t=1}^{2K} |\mathcal{Z}_t(\mathbf{x})| \cdot [P_0(t)]^M. \tag{5.8}$$

By allocating the failure budget as $\epsilon/(2K)$ per weight class and applying Lemma 5.6,

$$|\mathcal{Z}_t(\mathbf{x})|\cdot[P_0(t)]^M \;\leq\; \left(\frac{eN(q-1)}{t}\right)^t [P_0(t)]^M \;\leq\; \frac{\epsilon}{2K} \quad\Longleftarrow\quad M\,\log_2\frac{1}{P_0(t)} \;\geq\; t\log_2\frac{eN(q-1)}{t}+\log_2\frac{2K}{\epsilon},$$

which rearranges to the per-t condition

$$M \;\geq\; \frac{t\log_2(eN(q-1)/t)+\log_2(2K/\epsilon)}{\log_2(1/P_0(t))}, \qquad t = 1,\ldots,2K. \tag{5.9}$$

Taking the maximum over t gives (5.6).

Define $R(t) := \dfrac{\log_2|\mathcal{Z}_t(\mathbf{x})|+\log_2(2K/\epsilon)}{\log_2(1/P_0(t))}$ for $t \in [1, 2K]$.

Since $\mathbf{z} \in \mathcal{X}_K$ has $\|\mathbf{z}\|_0 \le K$ and $\|\mathbf{x}\|_0 = K$ (Assumption 3.7), an error vector of weight t decomposes into s outside-supp($\mathbf{x}$) positions and $t-s$ inside-supp($\mathbf{x}$) positions, with $s \in [\max(0, t-K), \min(t, K)]$. Hence,

$$|\mathcal{Z}_t(\mathbf{x})| \;\le\; \sum_{s=\max(0,t-K)}^{\min(t,K)} \binom{K}{t-s}\binom{N-K}{s}(q-1)^t.$$

In the sparse signal regime $N \gg K$, the dominant contribution to $\log_2 |\mathcal{Z}_t(\mathbf{x})|$ scales as $\Theta(t\log_2(N/K) + t\log_2(q-1))$ at leading order, and is non-decreasing in t on $[1, 2K]$.

Since $\alpha_t(d) = \binom{N-t}{d}/\binom{N}{d}$ is strictly decreasing in t, $\log_2(1/P_0(t)) = \log_2 q - \log_2(1 + \alpha_t(d)(q-1))$ is strictly increasing in t but bounded above by $\log_2 q$, which is a t-independent constant.

The numerator of $R(t)$ linearly grows in t at the leading order, while the denominator is bounded. Hence, at the leading order in the sparse regime $N \gg K$ with $d = o(N)$, $R(t)$ is non-decreasing in t on $[1, 2K]$ and

$$\max_{t\in[1,2K]} R(t) \;=\; R(2K) \quad \text{at the leading order.}$$

Substituting Lemma 5.6 at $t = 2K$ (specifically (5.5)) into (5.9) gives the following:

$$M \;\ge\; \frac{2K\log_2(eN(q-1)/(2K)) + \log_2(2K/\epsilon)}{\log_2(q/(1+\alpha_{2K}(d)(q-1)))},$$

which is (5.7). □

Remark 5.9 (Structural implications). As d grows in the sparse signal regime, $\alpha_{2K}(d) \approx e^{-2dK/N} \to 0$; thus, the denominator of (5.7) grows toward $\log_2 q$. This benefit saturates once $\alpha_{2K}(d) \lesssim 1/q$. The denominator grows as $\log_2 q$, while the dominant numerator term grows as $2K\log_2(q-1)$; thus, the sample complexity decreases with q at the leading order. The $\log_2(2K/\epsilon)$ term contributes an $O(\log K)$ correction; the dominant order is $\Theta(K\log_q(N/K))$.

Remark 5.10 (Order-optimality). Comparing Corollary 4.4 with Theorem 5.7, the asymptotic bounds coincide at $M = \Theta(K\log_q(N/K))$ when d is chosen so that $\alpha_{2K}(d)$ enters the saturation regime $\alpha_{2K}(d) \asymp 1/q$. The corresponding row weight $d^* \approx (N/(2K))\ln q$ is analyzed in Section 6, and the formal order-match is established in Theorem 6.4.

6. Saturation point of the row weight and gap analysis

As established in Sections 4 and 5, both the lower bound on M and the MAP-based achievability threshold on M depend on the row weight d. To compare them within a common framework, we introduce the following shorthand.

Definition 6.1 (Bound functionals in d)**.** For a fixed error tolerance $\epsilon \in (0, 1)$, define the following:

$$M_{LB}(d) := \frac{NH_2(K/N) + K\log_2(q-1)}{(1-\alpha_K(d))\log_2 q}, \tag{6.1}$$

$$M_{UB}(d) := \frac{2K\log_2(eN(q-1)/(2K)) + \log_2(2K/\epsilon)}{\log_2(q/(1+\alpha_{2K}(d)(q-1)))}. \tag{6.2}$$

By Corollary 4.4, any M that achieves $P_e \le \epsilon$ satisfies $M \ge M_{LB}(d)(1 - o(1)) - o(1)$ in the asymptotic regime; by Theorem 5.7, $M \ge M_{UB}(d)$ suffices.

The objective of this section is to identify the smallest d within the sparse pooling matrix regime $d = o(N)$ at which $M_{UB}(d)$ enters its leading-order saturation, and to compare the two functionals at that point. The restriction $d = o(N)$ is required both for the pooling matrix to remain sparse and for the multiplicative approximation $\alpha_t(d) = e^{-td/N}(1 + o(1))$ (valid as $K/N \to 0$ and $d = o(N)$ with $td^2/N^2 + t^2d/N^2 = o(1)$) to hold.

6.1. Saturation point d^*

The per-observation rate in the upper bound is

$$D_{UB}(d) \;:=\; \log_2\!\left(\frac{q}{1 + \alpha_{2K}(d)(q-1)}\right), \tag{6.3}$$

with $\alpha_{2K}(d) = e^{-2Kd/N}(1 + o(1))$ in the regime above. As d grows (within $d = o(N)$), $\alpha_{2K}(d)$ monotonically decreases toward 0; thus, $D_{UB}(d)$ monotonically increases toward $\log_2 q$. Consequently, $M_{UB}(d)$ is monotonically decreasing in d with diminishing returns: the additive penalty $\log_2(1 + \alpha_{2K}(d)(q-1))$ in the denominator becomes $O(1)$ once $\alpha_{2K}(d) \lesssim 1/q$.

Proposition 6.2 (Saturation point of the row weight)**.** *Define the* saturation point d^* *as the unique positive real solution of*

$$\alpha_{2K}(d) \;=\; \frac{1}{q}, \tag{6.4}$$

under the continuous extension $\alpha_{2K}(d) := e^{-2Kd/N}$. Then,

$$d^* \;=\; \frac{N}{2K}\ln q, \tag{6.5}$$

and in the discrete pooling problem, one selects $d = \lceil d^* \rceil$ (the effect on the leading-order analysis is $o(1)$). At $d = d^*$,

$$D_{UB}(d^*) \;=\; \log_2\!\left(\frac{q^2}{2q-1}\right) \;=\; \log_2 q - \log_2\!\left(2 - \tfrac{1}{q}\right), \tag{6.6}$$

which differs from $\log_2 q$ by an additive term in $[\log_2(3/2), 1]$ for all $q \geq 2$. Moreover, in the regime $K \to \infty$ with $K/N \to 0$ and $\log q = o(K)$, $d^* = \Theta((N/K)\log q) = o(N)$; therefore,$d^*$ lies in the sparse regime.

Proof. The equation $e^{-2Kd/N} = 1/q$ yields $d^* = (N/(2K))\ln q$. Substituting $\alpha_{2K}(d^*) = 1/q$ into (6.3) gives the following:

$$D_{UB}(d^*) = \log_2\frac{q}{1 + (1/q)(q-1)} = \log_2\frac{q}{(2q-1)/q} = \log_2\frac{q^2}{2q-1}.$$

The identity $\log_2(q^2/(2q-1)) = \log_2 q - \log_2(2 - 1/q)$ is algebraic, and $\log_2(2 - 1/q) \in [\log_2(3/2), 1]$ for $q \geq 2$. The condition $d^* = o(N)$ is equivalent to $(\log q)/K \to 0$ ($\log q = o(K)$). □

Remark 6.3 (Saturation versus unconstrained minimization). d^* is the saturation point of $D_{UB}(d)$, not an interior minimizer of $M_{UB}(d)$. Since the numerator of $M_{UB}(d)$ in (6.2) is independent of d and the denominator $D_{UB}(d)$ is strictly increasing in d, $M_{UB}(d)$ has no interior minimizer; its infimum over

$d \in [1, N]$ is only approached as $d \to N$, at which point the sparse-regime assumption $d = o(N)$ and the exponential approximation both fail. Hence, d^* is the smallest row weight at which $D_{UB}(d)$ reaches its leading-order limit $\log_2 q$ within the sparse regime. Row weights beyond d^* only yield sub-leading improvements and are not order-relevant.

6.2. *Order-match of the two bounds at $d = d^*$*

Now, We compare $M_{LB}(d^*)$ and $M_{UB}(d^*)$ at the leading order.

Theorem 6.4 (Order-matched bounds at d^*)**.** *Consider the regime $N \to \infty$ with $K \to \infty$, $K/N \to 0$, $N/K \to \infty$, $\log q = o(K)$, and $\epsilon \to 0$ slowly enough that $\log_2(2K/\epsilon) = o(K \log_2(N/K))$. Then, at $d = d^*$ defined in* (6.5)*,*

$$M_{LB}(d^*) \;=\; \Theta\Big(K \log_q(N/K)\Big), \qquad M_{UB}(d^*) \;=\; \Theta\Big(K \log_q(N/K)\Big), \tag{6.7}$$

and

$$\limsup_{N,K\to\infty} \frac{M_{UB}(d^*)}{M_{LB}(d^*)} \;\le\; C(q), \tag{6.8}$$

for a constant $C(q)$ which only depends on q, with $C(q) \to 2$ as $q \to \infty$.

Proof. At $d = d^*$, $\alpha_{2K}(d^*) = 1/q$ by (6.4), and

$$\alpha_K(d^*) = e^{-Kd^*/N}(1 + o(1)) = e^{-(\ln q)/2}(1 + o(1)) = q^{-1/2}(1 + o(1)).$$

Using $H_2(K/N) = (K/N)\log_2(eN/K) + O((K/N)^2)$ for $K/N \to 0$,

$$NH_2(K/N) = K \log_2(eN/K) + O(K^2/N).$$

The lower-bound numerator is

$$K \log_2(eN/K) + K \log_2(q-1) + O(K^2/N) = K \log_2(eN(q-1)/K)(1 + o(1)).$$

The upper-bound numerator is

$$2K \log_2(eN(q-1)/(2K)) + \log_2(2K/\epsilon) = 2K \log_2(eN(q-1)/K) - 2K + o(K \log_2(N/K)),$$

where the $\log_2(2K/\epsilon)$ term is $o(K \log_2(N/K))$ by hypothesis. The ratio of the two numerators tends to 2 as $N/K \to \infty$.

The lower-bound denominator is

$$(1 - \alpha_K(d^*)) \log_2 q = (1 - q^{-1/2}) \log_2 q + o(\log_2 q),$$

while the upper-bound denominator is

$$\log_2(q^2/(2q-1)) = \log_2 q - \log_2(2 - 1/q).$$

Both are $\Theta(\log_2 q)$. Their ratio

$$\rho(q) \;:=\; \frac{\log_2(q^2/(2q-1))}{(1 - q^{-1/2}) \log_2 q} \tag{6.9}$$

satisfies $\rho(q) \in (0, \infty)$ for every $q \geq 2$ and $\rho(q) \to 1$ as $q \to \infty$ (the numerator equals $\log_2 q - 1 + o(1)$ and the denominator equals $\log_2 q + o(\log_2 q)$).

Both $M_{LB}(d^*)$ and $M_{UB}(d^*)$ equal a ratio of numerator $\Theta(K \log_2(N/K))$ to denominator $\Theta(\log_2 q)$; hence, both are $\Theta(K \log_2(N/K)/ \log_2 q) = \Theta(K \log_q(N/K))$. The additive $\Theta(K)$ from the $K \log_2(q-1)$ term contributes $\Theta(K/ \log_2 q) = o(K \log_q(N/K))$ under $N/K \to \infty$ and is sub-leading. Taking the following ratio:

$$\frac{M_{UB}(d^*)}{M_{LB}(d^*)} = 2\rho(q)\,(1 + o(1));$$

thus, $C(q) = 2\rho(q)$. As $q \to \infty$, $\rho(q) \to 1$ and $C(q) \to 2$. □

Remark 6.5 (Structural interpretation of the gap). In the sparse signal regime, $NH_2(K/N) = K \log_2(eN/K)(1 + o(1))$, so the two numerators agree in order; the upper-bound numerator is larger by a factor that approaches 2, which is traceable to the $t = 2K$ bottleneck in the union bound of Theorem 5.7. The denominators differ because the lower bound depends on $\alpha_K(d)$ (a pool missing the state support) while the upper bound depends on $\alpha_{2K}(d)$ (a pool missing the worst-case error support). For a small d, the gap between $\alpha_K(d)$ and $\alpha_{2K}(d)$ is order-relevant. Setting $\alpha_{2K}(d^*) = 1/q$ brings the upper-bound denominator to $\Theta(\log_2 q)$; the lower-bound denominator at this d is $(1 - q^{-1/2}) \log_2 q$, also $\Theta(\log_2 q)$. Therefore, the order-level discrepancy in the denominator vanishes at d^*. Since $\log_2(q-1) = \log_2 q + o(1)$ as $q \to \infty$, the numerator-to-denominator ratio per active coordinate approaches 1; thus, a larger q reduces M in both bounds at the same order.

Choosing d at the saturation point $d^* = (N/(2K)) \ln q$ removes the leading-order discrepancy between the Fano necessary condition and the MAP sufficient condition. Theorem 6.4 shows that the two bounds match to within a constant $C(q) \leq 2\rho(q)$ which only depends on q, with $C(q) \to 2$ as $q \to \infty$. Therefore, the analysis yields an order-optimal sample complexity $\Theta(K \log_q(N/K))$ for the finite-field community state inference under the constant row weight ensembles.

7. Numerical evaluation of the analytical bounds

The asymptotic gap ratio established in Theorem 6.4 is a statement about the analytical bounds M_{LB} and M_{UB} derived in Sections 4 and 5. In this section, we numerically evaluate these bounds in order to visualize their dependence on the field size q and the relative sparsity ratio K/N, to inspect the convergence of the multiplicative gap toward 2 for large q, and to illustrate the role of $d^* = (N/(2K)) \ln q$ as the evaluation point in our analytical expressions.

We emphasize that the curves shown below are deterministic evaluations of the analytical bounds themselves; they are not the result of Monte-Carlo experiments with a specific decoder. As discussed in Section 2.4, exact MAP decoding has search-space cardinality $|\mathcal{X}_K| = \sum_{k=0}^{K} \binom{N}{k}(q-1)^k$, which is computationally intractable at the scale $N = 1000$ considered here, so a direct empirical validation of the upper bound is not within reach of a standard simulation. Plotting the analytical bounds yields a numerical illustration of the trends predicted by our theory while remaining within the information-theoretic scope of the paper.

All evaluations use the parameter setting $N = 1000$, $\epsilon = 10^{-3}$, and $q \in \{2, 4, 8, 16, 32, 64\}$, with K/N swept over $\{0.01, 0.02, \ldots, 0.15\}$ in Figures 1 and 2, and with $K/N = 0.05$ fixed in Figure 3. Throughout, M_{LB} is computed from Theorem 4.3 and M_{UB} from the sparse-regime specialization at

$t = 2K$ derived in Section 5.2. Combinatorial quantities are evaluated in logarithmic form via the log-gamma function in order to maintain numerical stability at $N = 1000$; in particular, $\alpha_t(d) = \binom{N-t}{d}/\binom{N}{d}$ in Lemma 5.4 is computed as $\exp(\ln\binom{N-t}{d} - \ln\binom{N}{d})$.

7.1. Scaling of the bounds with K/N

Figure 1 reports M_{LB} and M_{UB} as a function of K/N for the six field sizes under consideration, with d fixed at the saturation point d^*. For every fixed q, both bounds are monotonically increasing in K/N, which is consistent with the observation that recovering a larger support requires more observations. For every fixed K/N, both bounds are monotonically decreasing in q: larger alphabets carry more information per observation, captured by the $\log_2 q$ factor in the denominators of both expressions. The two bounds exhibit similar scaling in K/N—their trajectories on the linear plots are parallel—with the multiplicative gap controlled by q alone. This parallelism is the finite-N manifestation of the asymptotic relation $M_{UB} \sim 2\rho(q)\, M_{LB}$ established in Theorem 6.4.

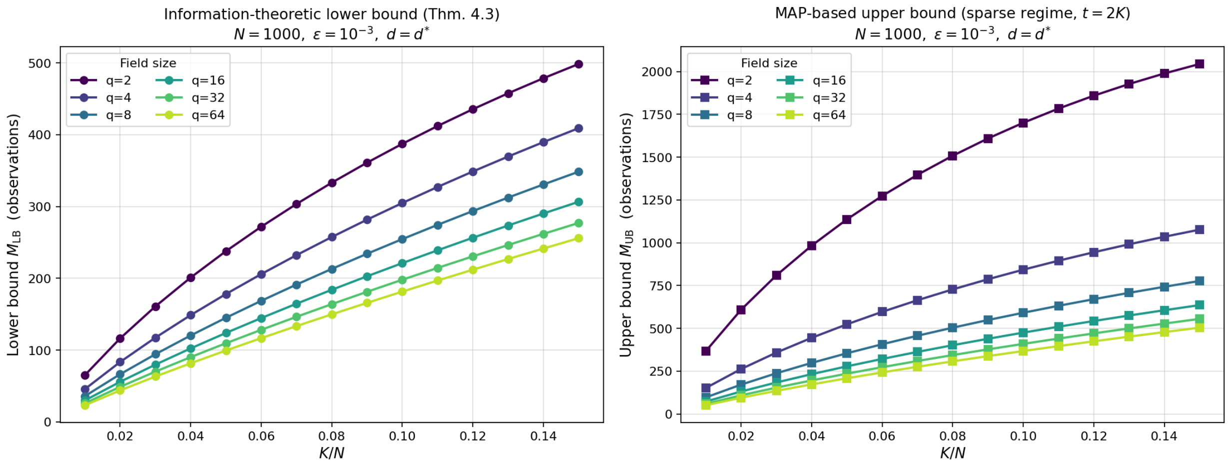

Figure 1. Evaluation of the analytical bounds for $N = 1000$, $\epsilon = 10^{-3}$, $d = d^*$, and $q \in \{2, 4, 8, 16, 32, 64\}$. Left: information-theoretic lower bound M_{LB} from Theorem 4.3. Right: upper bound M_{UB} in the sparse signal regime ($t = 2K$ specialization). These are direct evaluations of the analytical bound expressions and are not Monte-Carlo results.

7.2. Multiplicative gap and convergence toward 2

To examine the asymptotic statement $C(q) \to 2$ at finite N, Figure 2 plots the ratio M_{UB}/M_{LB} as a function of K/N together with a horizontal reference line at the asymptote 2. At $q = 64$, the ratio remains below 2.2 across the entire K/N range and dips slightly below 2 at the smallest K/N; this sub-asymptote excursion is a finite-N artifact tied to the $\binom{N-2K}{d^*}/\binom{N}{d^*}$ approximation and is consistent with the asymptotic statement. At the opposite extreme, for $q = 2$, the ratio ranges from approximately 5.7 at a small K/N down to about 4.1 at $K/N = 0.15$, above the asymptote; this is consistent with the fact that $\rho(q) \to 1$ only as $q \to \infty$, and the constant $2\rho(2)$ predicted by Theorem 6.4 is larger than 2. Across all values of q, the ratio is monotonically decreasing in K/N within this range, thus indicating that the bound pair becomes tightest for a larger K at a fixed N. These observations provide a finite-N, finite-q inspection of the asymptotic gap ratio without serving as a proof of the asymptotic theorem itself.

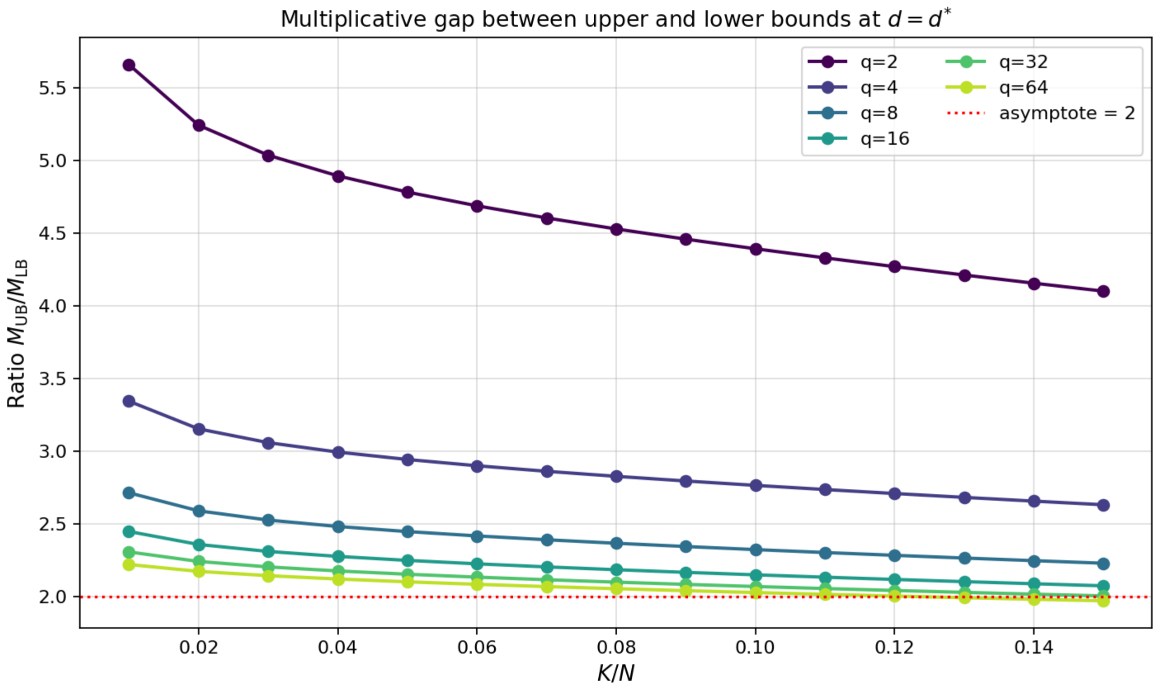

Figure 2. Multiplicative gap M_{UB}/M_{LB} between the analytical bounds, evaluated at $d = d^*$. The dotted line at 2 is the asymptote established in Theorem 6.4. For $q = 64$, the finite-N ratio is within a few percent of 2; for $q = 2$, a larger gap reflects the fact that $\rho(q) \to 1$ only as $q \to \infty$.

7.3. *Role of the saturation point d^**

A natural question is whether $d = d^*$ is the row weight that minimizes $M_{UB}(d)$. It is not. As shown in Figure 3, the map $d \mapsto M_{UB}(d)$ is strictly decreasing on its natural domain, because $\alpha_{2K}(d)$ is strictly decreasing in d and the denominator $\log_2(q/(1+\alpha_{2K}(d)(q-1)))$ is therefore strictly increasing in d. The role of d^* is that of a saturation point (cf. Remark 6.3): it is the value at which $\alpha_{2K}(d) \approx 1/q$, beyond which the marginal reduction in $M_{UB}(d)$ is small relative to the reduction achieved on the interval $[1, d^*]$. For $q = 4$ and the parameters above, the bound at $d = d^* \approx 13.86$ is within a few percent of its infimum as $d \to N - 2K$. Therefore, choosing $d = d^*$ captures most of the achievable reduction in M_{UB} while keeping the analytical expressions tractable and the dependence on q explicit. This is the sense in which our bound expressions, and the asymptotic constant $C(q)$ in Theorem 6.4, are stated at $d = d^*$: the evaluation point is selected not for strict minimization, but for the combination of analytical tractability and near-saturation that the figure illustrates.

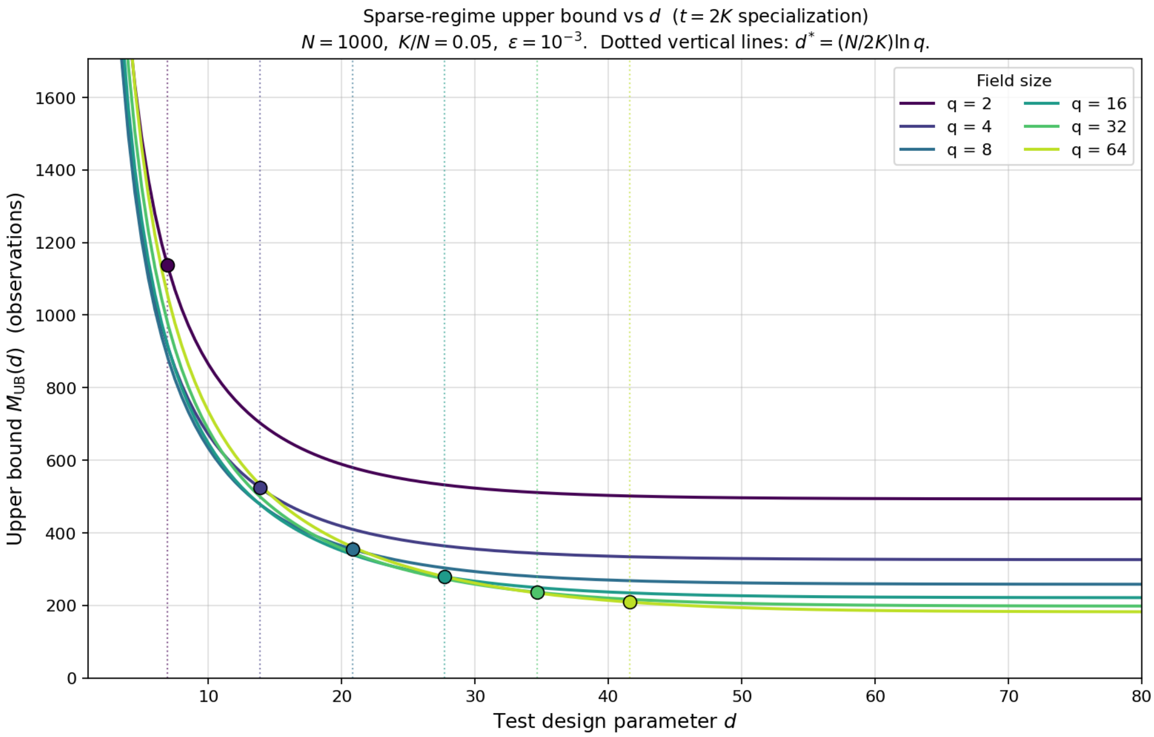

Figure 3. Sparse-regime upper bound $M_{UB}(d)$ as a function of the constant row weight d, for $N = 1000$, $K/N = 0.05$, and $\epsilon = 10^{-3}$. Dotted vertical lines indicate $d^* = (N/(2K)) \ln q$ for each q; circular markers report the corresponding bound value. The map $d \mapsto M_{UB}(d)$ is strictly decreasing; the marginal reduction beyond d^* is small relative to that on $[1, d^*]$, so the saturation point captures most of the achievable reduction.

8. Conclusions

We studied the sample complexity limits of finite-field community state inference, formulated as the inference of a K-sparse state vector $\mathbf{x} \in \mathbb{F}_q^N$ from M pooled observations $\mathbf{y} = \mathbf{A}\mathbf{x}$ over $\mathbb{F}_q$, where $\mathbf{A} \in \{0, 1\}^{M \times N}$ is a binary pooling matrix with a constant row weight of d. Fano's inequality (Theorem 4.3) gives the necessary condition on M, and an union bound over $2K$-sparse error vectors (Theorem 5.7) gives a sufficient condition.

The main result is an order-level match between the two bounds at the saturation point $d^* = (N/(2K)) \ln q$, defined in Proposition 6.2 as the unique positive real solution of $\alpha_{2K}(d) = 1/q$ under the continuous extension. Under the asymptotic regime $K \to \infty$, $K/N \to 0$, $N/K \to \infty$, and $\log q = o(K)$, Theorem 6.4 shows that both bounds are of order $\Theta(K \log_q(N/K))$ and that their ratio is asymptotically bounded by a q-dependent constant tending to 2 as $q \to \infty$. We note that d^* is the saturation point of the per-observation rate within the sparse pooling matrix regime $d = o(N)$, not an interior minimizer of the upper bound; the deterministic evaluations of Section 7 illustrate this distinction at a finite N.

This order-match shows that random pooling matrices with a constant row weight attain $\Theta(K \log_q(N/K))$ samples under the noiseless model considered here. We do not claim that they are capacity-achieving in the algebraic-coding sense, nor that the constant-factor gap—traceable to the $t = 2K$ bottleneck in the union bound—is tight; tightening it would require a refined achievability analysis or a strengthened converse.

The bounds characterize the sample complexity that is necessary and sufficient, not how it is attained by a polynomial-time decoder. Closing this algorithmic gap is the natural next step, through a density-evolution analysis of non-binary belief propagation on the constant row weight ensembles considered here, in the spirit of [29], through information-set decoding [31, 33] adapted to the $\mathbb{F}_q$-sparse setting, or through structured constant row weight parity-check matrices that admit Berlekamp–Massey-type decoders. A Monte-Carlo evaluation of the gap between the information-theoretic limit and concrete practical decoders, as well as extensions to noisy observations and to richer community-structure models (hierarchical, overlapping, or heterogeneous-degree), are left for future work.

Use of Generative-AI tools declaration

The author declares he has not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528).

Conflict of interest

The author declares no conflicts of interest.

References

1. M. Aldridge, O. Johnson, J. Scarlett, Group testing: An information theory perspective, *Found. Trends Commun.*, **15** (2019), 196–392. https://doi.org/10.1561/0100000099
2. P. Nikolopoulos, S. R. Srinivasavaradhan, T. Guo, C. Fragouli, S. Diggavi, Group testing for overlapping communities, *ICC 2021–IEEE International Conference on Communications*, Montreal, QC, Canada, 2021, 1–7. https://doi.org/10.1109/ICC42927.2021.9500791
3. S. Ahn, W.-N. Chen, A. Özgür, Adaptive group testing on networks with community structure: The stochastic block model, *IEEE T. Inform. Theory*, **69** (2023), 4758–4776, https://doi.org/10.1109/TIT.2023.3247520
4. S. Fortunato, Community detection in graphs, *Physics Reports*, **486** (2010), 75–174, https://doi.org/10.1016/j.physrep.2009.11.002
5. B. Karrer, M. E. J. Newman, Stochastic blockmodels and community structure in networks, *Phys. Rev. E*, **83** (2011), 016107, https://doi.org/10.1103/PhysRevE.83.016107
6. E. Abbe, Community detection and stochastic block models: recent developments, *J. Mach. Learn. Res.*, **18** (2018), 6446–6531.
7. D. L. Donoho, Compressed sensing, *IEEE T. Inform. Theory*, **52** (2006), 1289–1306. https://doi.org/10.1109/TIT.2006.871582

8. E. J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE T. Inform. Theory*, **52** (2006), 489–509. https://doi.org/10.1109/TIT.2005.862083
9. D.-Z. Du, F. K. Hwang, *Combinatorial group testing and its applications*, 2 Eds., Singapore: World Scientific, 2000. https://doi.org/10.1142/4252
10. J.-T. Seong, H.-N. Lee, Necessary and sufficient conditions for recovery of sparse signals over finite fields, *IEEE Commun. Lett.*, **17** (2013), 1976–1979. https://doi.org/10.1109/LCOMM.2013.090313.130753
11. A. K. Das, S. Vishwanath, On finite-alphabet compressive sensing, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013, 5890–5894. https://doi.org/10.1109/ICASSP.2013.6638794
12. R. Ahlswede, N. Cai, S.-Y. R. Li, R. W. Yeung, Network information flow, *IEEE T. Inform. Theory*, **46** (2000), 1204–1216. https://doi.org/10.1109/18.850663
13. S.-Y. R. Li, R. W. Yeung, N. Cai, Linear network coding, *IEEE T. Inform. Theory*, **49** (2003), 371–381. https://doi.org/10.1109/TIT.2002.807285
14. G. K. Atia, V. Saligrama, Boolean compressed sensing and noisy group testing, *IEEE T. Inform. Theory*, **58** (2012), 1880–1901. https://doi.org/10.1109/TIT.2011.2178156
15. A. Mazumdar, Nonadaptive group testing with random set of defectives, *IEEE T. Inform. Theory*, **62** (2016), 7522–7531. https://doi.org/10.1109/TIT.2016.2613870
16. T. M. Cover, J. A. Thomas, *Elements of information theory*, Hoboken: John Wiley & Sons, 2005. https://doi.org/10.1002/047174882X
17. M. E. J. Newman, Modularity and community structure in networks, *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006), 8577–8582. https://doi.org/10.1073/pnas.0601602103
18. V. A. Traag, L. Waltman, N. J. van Eck, From louvain to leiden: Guaranteeing well-connected communities, *Sci. Rep.*, **9** (2019), 5233. https://doi.org/10.1038/s41598-019-41695-z
19. Z. X. Zhou, A. A. Amini, Analysis of spectral clustering algorithms for community detection: the general bipartite setting, *J. Mach. Learn. Res.*, **20** (2019), 1774–1820.
20. H. Qing, J. L. Wang, An improved spectral clustering method for community detection under the degree-corrected stochastic blockmodel, 2010, arXiv:2011.06374. https://doi.org/10.48550/arXiv.2011.06374
21. X. Y. Lu, B. K. Szymanski, A regularized stochastic block model for robust community detection in complex networks, *Sci. Rep.*, **9** (2019), 13247. https://doi.org/10.1038/s41598-019-49580-5
22. V. Cohen-Addad, F. Mallmann-Trenn, D. Saulpic, Community recovery in the degree-heterogeneous stochastic block model, *Proceedings of Thirty Fifth Conference on Learning Theory*, PMLR, 2022, 1662–1692.
23. M.-J. Lai, D. Mckenzie, A compressive sensing approach to community detection with applications, 2018, arXiv:1708.09477. https://doi.org/10.48550/arXiv.1708.09477
24. T. J. Wang, J. B. Tan, W. B. Ding, Y. R. Zhang, F. Yang, J. Song, et al., Intercommunity detection scheme for social internet of things: compressive sensing over graphs approach, *IEEE Internet Things*, **5** (2018), 4550–4557. https://doi.org/10.1109/JIOT.2018.2837048

25. P. Nikolopoulos, S. R. Srinivasavaradhan, T. Guo, C. Fragouli, S. N. Diggavi, Community-aware group testing, *IEEE T. Inform. Theory*, **69** (2023), 4361–4383. https://doi.org/10.1109/TIT.2023.3250119
26. R. M. Roth, *Introduction to coding theory*, Cambridge: Cambridge University Press, 2006, https://doi.org/10.1017/CBO9780511808968
27. D. H. Wiedemann, Solving sparse linear equations over finite fields, *IEEE T. Inform. Theory*, **32** (1986), 54–62. https://doi.org/10.1109/TIT.1986.1057137
28. M. C. Davey, D. J. C. MacKay, Low-density parity check codes over GF(q), *IEEE Commun. Lett.*, **2** (1998), 165–167. https://doi.org/10.1109/4234.681360
29. D. Declercq, M. Fossorier, Decoding algorithms for nonbinary LDPC codes over GF(q), *IEEE T. Commun.*, **55** (2007), 633–643. https://doi.org/10.1109/TCOMM.2007.894088
30. E. R. Berlekamp, *Algebraic coding theory*, New York: McGraw-Hill, 1968.
31. J. Stern, A method for finding codewords of small weight, In: *Coding theory and applications*, Berlin: Springer, 1989, 106–113. https://doi.org/10.1007/BFb0019850
32. E. R. Berlekamp, R. J. McEliece, H. C. A. van Tilborg, On the inherent intractability of certain coding problems, *IEEE T. Inform. Theory*, **24** (1978), 384–386. https://doi.org/10.1109/TIT.1978.1055873
33. C. Peters, Information-set decoding for linear codes over $\mathbb{F}_q$, In: *Post-quantum cryptography*, Berlin: Springer, 2010, 81–94. https://doi.org/10.1007/978-3-642-12929-2_7
34. D. L. Donoho, A. Maleki, A. Montanari, Message-passing algorithms for compressed sensing, *Proc. Natl. Acad. Sci. U.S.A.*, **106** (2009), 18914–18919. https://doi.org/10.1073/pnas.0909892106
35. S. Rangan, Generalized approximate message passing for estimation with random linear mixing, *2011 IEEE International Symposium on Information Theory Proceedings*, St. Petersburg, Russia, 2011, 2168–2172. https://doi.org/10.1109/ISIT.2011.6033942
36. A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, P. Loick, Optimal group testing, In: *Proceedings of Thirty Third Conference on Learning Theory* Learning Research, PMLR, 2020, 1374–1388.
37. M. Abramowitz, I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Washington: National Bureau of Standards, 1964.
38. J. Scarlett, V. Cevher, Phase transitions in group testing, *The Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, Arlington, Virginia, 2016, 40–53.