

Improved bounds on the size of sparse parity check matrices

extended abstract

Assaf Naor
Microsoft Research
anaor@microsoft.com

Jacques Verstraëte
University of Waterloo
jverstra@fe01.math.uwaterloo.ca

Abstract

Let $N_{\mathbb{F}}(n, k, r)$ denote the maximum number of columns in an n -row matrix with entries in a finite field \mathbb{F} in which each column has at most r nonzero entries and every k columns are linearly independent over \mathbb{F} . Such sparse parity check matrices are fundamental tools in coding theory, derandomization and complexity theory. We obtain near-optimal theoretical upper bounds for $N_{\mathbb{F}}(n, k, r)$ in the important case $k > r$, i.e. when the number of correctable errors is greater than the weight. Namely, we show that $N_{\mathbb{F}}(n, k, r) = O(n^{\frac{r}{2} + \frac{4r}{3k}})$. The best known (probabilistic) lower bound [20] is $N_{\mathbb{F}}(n, k, r) = \Omega(n^{\frac{r}{2} + \frac{r}{2k-2}})$, while the best known upper bound [20] in the case $k > r$ was for k a power of 2, in which case $N_{\mathbb{F}}(n, k, r) = \Omega(n^{\frac{r}{2} + \frac{1}{2}})$. Our method is based on a novel reduction of the problem to the extremal problem for cycles in graphs, and yields a fast algorithm for finding short linear dependences in large sets of sparse vectors. In the full version of this paper we present additional applications of this method to problems in combinatorial number theory.

1 Introduction

Low density parity check codes were introduced by Gallager in the 1960s, and have since found numerous theoretical and practical applications in computer science (see [16, 22, 23] for an account of this theory. We also refer to [11] for a nice introduction to the geometry of codes). Given a linear code $C \subseteq \mathbb{F}_2^m$ of dimension ℓ and minimum Hamming weight t , an $(m - \ell) \times m$ matrix H is called a *parity check matrix* of C if $C = \{v \in \mathbb{F}_2^m : Hv = 0\}$. We shall say that H is r -sparse if every column of H has at most r non-zero entries. The *Syndrome Decoding Algorithm* for such codes works as follows: given a corrupted signal z one computes the vector x of minimal weight satisfying $Hx = Hz$, and decodes z to $z - x$ (this algorithm corrects at most t errors). As such computations are faster if the sparseness of H is exploited, it is desirable to obtain codes with sparse parity check matrices. Indeed, sparse parity check matrices occur in many of the known constructions of codes, e.g. codes based on bounded degree graphs such as expander codes [25, 26], and we also refer to [21] for theoretical and experimental coding theory applications of very sparse matrices (we stress here that the present paper deals with a different range of parameters – our bounds will be for codes in which the minimal weight is not proportional to the dimension. Such codes occur in several contexts, e.g. certain BCH and Reed-Solomon codes [22], Turbo and Turbo-like codes [8, 18, 13, 7]). Additionally, the above discussion makes sense for parity check matrices over arbitrary finite fields, which are also used in coding theory (see [22, 23] for basic information on this topic, and [14] for empirical results on such codes). Finally, sparse parity check matrices are the key ingredient in the construction of small probability spaces and deterministic simulations of k -wise independent random variables, which are a key tool in derandomization [1, 2, 4, 10].

Somewhat surprisingly, in spite of their importance, there was a large gap between the known upper and lower bounds for the maximal number of columns of sparse parity check matrices. For a finite field \mathbb{F} let $N_{\mathbb{F}}(n, k, r)$ be the maximal number of vectors in \mathbb{F}^n with at most r non-zero coordinates such that no k of them are linearly dependent (observe that the linear independence condition corresponds to the fact that the kernel of the matrix whose rows are the given vectors is a code with minimal distance at least $k + 1$). When $\mathbb{F} = \mathbb{F}_2$ we use the notation $N_{\mathbb{F}_2}(n, k, r) = N(n, k, r)$. The problem dealt with in this paper, namely that of estimating $N_{\mathbb{F}}(n, k, r)$, differs from the classical Gilbert-Varshamov bounds [22], since the classical bounds on sizes of codes are geometric packing bounds which depend only on the minimum distance of the code. Here we introduce an additional algebraic restriction on the code (the existence of a sparse parity check matrix) which is motivated by computational issues. Thus, we are dealing with a mixture of a geometric and algebraic problem. In this paper we deal with the case of k, r fixed and $n \rightarrow \infty$ (actually, our bounds improve the known bounds up to $k = n^{o(1)}$). A probabilistic construction [20] (using the first moment method) shows that

$$N(n, k, r) = \Omega\left(n^{\frac{r}{2} + \frac{r}{2k-2}}\right). \quad (1)$$

and this was generalized to arbitrary finite fields in [19] to $N_{\mathbb{F}}(n, k, r) = \Omega\left(n^{\frac{r}{2} + \frac{r}{2k-2}}\right)$ for even k and $N_{\mathbb{F}}(n, k, r) = \Omega\left(n^{\frac{r}{2} + \frac{r}{2k-4}}\right)$ for odd k . When $k \geq 4$ is even, and $\gcd(k - 1, r) = 1$, the lower bound (1) was improved in [9] to

$$N(n, k, r) = \Omega\left(n^{\frac{r}{2} + \frac{r}{2k-2}} \cdot (\log n)^{\frac{1}{k-1}}\right).$$

This lower bound was generalized to arbitrary finite fields in [19] (in which case the constant also depends on the size of the field).

In [20] it was shown that when k is a power of 2, $N(n, k, r) = O(n^{\frac{1}{2}\lceil r + \frac{r}{k-1} \rceil})$, which coincides with the lower bound (1) (up to factors independent of n) when $k-1$ divides r . This upper bound was generalized to arbitrary finite fields in [19]. Observe that in the important case $k > r$, i.e. when the number of correctable errors is greater than the weight, these upper bounds become: for k a power of 2, $N_{\mathbb{F}}(n, k, r) = O(n^{\frac{r}{2} + \frac{1}{2}})$. Thus the gap between the exponent of n in this bound and the probabilistic lower bounds deteriorates as k grows. The main theorem in this paper is the following:

Theorem 1.1. *For every k and every finite field \mathbb{F}*

$$N_{\mathbb{F}}(n, k, r) = O(n^{\frac{r}{2} + \frac{4r}{3k}}),$$

where the implied constant depends only on k, r and $|\mathbb{F}|$.

In particular it follows from Theorem 1.1 that for any positive integer r ,

$$\lim_{k \rightarrow \infty} \liminf_{n \rightarrow \infty} \frac{\log N_{\mathbb{F}}(n, k, r)}{\log n} = \lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{\log N_{\mathbb{F}}(n, k, r)}{\log n} = \frac{r}{2}.$$

It is worthwhile to note here that the proof of Theorem 1.1 when r is even differs markedly from the proof in the case of odd r . In fact, it turns out that the case of odd r involves a substantially more subtle argument. The difference between these cases will be explained in section 2. It seems likely that evaluating the smallest $c = c(k, r)$ for which $N(n, k, r) = O(n^{\frac{r}{2} + \frac{cr}{k} + \epsilon})$ for all $\epsilon > 0$ is a difficult problem (the answer probably depends on arithmetic properties of k and r). Our proof of Theorem 1.1 yields an algorithm which, given a set $X \subseteq \mathbb{F}^n$ of n -dimensional vectors of weight at most r satisfying $|X| = \Omega(n^{\frac{r}{2} + \frac{r}{k}})$, finds k linearly dependent vectors in X in time $O(|X|^4) = O(n^{4r})$ (observe that a trivial check of all possible $\binom{|X|}{k}$ linear dependencies requires time $\Omega(n^{\frac{kr}{2}})$).

The proof of Theorem 1.1 is based on a novel reduction of the problem to the following Turán type problem: What is the maximal number of edges in an n -vertex graph which doesn't contain an even cycle of length $2k$? We then employ recent results on this problem [27, 3, 17, 24] to deduce bounds on $N_{\mathbb{F}}(n, k, r)$. Some of the previous results on $N_{\mathbb{F}}(n, k, r)$ reduced the problem to the study of certain Turán type questions on hypergraphs (see [9, 7]). Since very little is known on hypergraph Turán problems, our main contribution is the method of reducing such questions to a problem on graphs. We believe that this approach is of independent interest. Indeed, in the full version of this paper we apply our method to problems in combinatorial number theory, improving in particular results of Erdős, Saárközy and Sós [15].

2 Forbidden cycles and sparse parity check matrices

Throughout this section, \mathbb{F} is a finite field and $\mathbb{F}^* = \mathbb{F} \setminus \{0\}$ is the set of non-zero elements of \mathbb{F} . A set $X \subseteq \mathbb{F}^n$ is *k -wise independent* if no k vectors in X are linearly dependent. A vector $v \in \mathbb{F}^n$ is said to have weight r if it has exactly r non-zero coordinates. Then $N_{\mathbb{F}}(n, k, r)$ is the maximum size of a set of k -wise independent vectors of weight at most r in \mathbb{F}^n .

In what follows, given two graphs $G = (V, E)$, $G' = (V', E')$ we write $G' \subseteq G$ if G contains a (not necessarily induced) copy of G' . The *girth* of a graph $G = (V, E)$ is the length of the shortest cycle in G . The following results on forbidden cycles will be used in the proof of Theorem 1.1:

Theorem 2.1. *Let k be a positive integer, and denote by C_{2k} the cycle of length $2k$.*

1. *If $G = (V, E)$ is an n -vertex graph such that $C_{2k} \not\subseteq G$, then $|E| < 8(k-1)n^{1+\frac{1}{k}}$.*
2. *If $G = (V, E)$ is an M by N bipartite graph such that $C_{2k} \not\subseteq G$, then*

$$|E| < \begin{cases} 2k \left[(MN)^{\frac{1}{2} + \frac{1}{2k}} + M + N \right] & \text{if } k \text{ is odd,} \\ 2k \left[MN^{\frac{1}{2} + \frac{1}{k}} + M + N \right] & \text{if } k \text{ is even.} \end{cases}$$

3. *If $G = (V, E)$ is a graph of girth $2k + 1$ then $|E| \leq \frac{1}{2}n^{1+\frac{1}{k}} + n$.*

The first assertion is proved in [27] (the same dependence on n , with a worse constant, is Erdős' *Even Cycle Theorem* – see [12]). The second assertion is proved in [24] and the third assertion is proved in [3] (see also [17] for a bipartite version of this result).

We will reduce the problem of estimating $N_{\mathbb{F}}(n, k, r)$ to the bounds in Theorem 2.1. The reduction we give is simple in the case that r is even, but substantially more involved when r is odd. The first theorem we prove is as follows:

Theorem 2.2. *Let n, k, r be positive integers, and define*

$$M = \sum_{\ell=0}^{\lfloor r/2 \rfloor} (|\mathbb{F}| - 1)^\ell \binom{n}{\ell} \quad \text{and} \quad N = \sum_{\ell=0}^{\lfloor r/2 \rfloor} (|\mathbb{F}| - 1)^\ell \binom{n}{\ell}.$$

Let $X \subseteq \mathbb{F}^n$ be a set of vectors of weight at most r such that

$$|X| > \begin{cases} 8(k-1)M^{1+\frac{1}{k}} & \text{if } r \text{ is even.} \\ 2k \left((MN)^{\frac{1}{2} + \frac{1}{2k}} + M + N \right) & \text{if } r, k \text{ are odd} \\ 2k \left(N^{\frac{1}{2}} M^{\frac{1}{2} + \frac{1}{k}} + M + N \right) & \text{if } r \text{ is odd and } k \text{ is even.} \end{cases}$$

Then there are disjoint sets $A, B \subseteq X$, each of size k , such that $\sum_{a \in A} a = \sum_{b \in B} b$.

The bounds in Theorem 2.2 show that $N_{\mathbb{F}}(n, 2k, r) = O(n^{\frac{r+1}{2}})$ for all k . This generalizes the same bound which was proved for k a power of two in [20] to all values of k .

When $k > r$ the first assertion of Theorem 2.2 implies Theorem 1.1 in the case of even r . In the case of odd r , Theorem 2.2 is insufficient to deduce Theorem 1.1, since the rounding up of $\frac{r}{2}$ only yields an upper bound of $O(n^{\frac{r}{2} + \frac{1}{2}})$. This fundamental difference between even and odd values of r leads to a more involved argument in the case of odd r , which nevertheless yields the bounds of Theorem 1.1 in this case as well. In fact, the ideas and constructions required for the case of odd r are crucial for our applications to combinatorial number theory. As the proof of the odd case is beyond the scope of this extended abstract, we defer it to the full version of this paper, and present a complete proof of the much simpler case of even r (which nevertheless contains some of the key ideas of the proof of the full result).

Proof of Theorem 2.2. For each vector $v \in X$ of weight $\omega \leq r$, fix a pair $e(v) = \{x, y\}$ of vectors of weights at most $\lfloor \omega/2 \rfloor$ and $\lceil \omega/2 \rceil$, respectively, in \mathbb{F}^n , satisfying $x + y = v$. If r is even, let G be the graph whose vertex set consists of all vectors in \mathbb{F}^n of weight at most $r/2$ and whose edge set is $E = \{e(v) : v \in X\}$. Then $|E| = |X|$ and G has M vertices. By the first assertion in Theorem 2.1, $C_{2k} \subseteq G$. By the definition of G , there exist distinct vectors $v_1, v_2, \dots, v_{2k} \in X$ such that the edge set of this cycle consist of the pairs

$$e(v_i) = \{x_i, x_{i+1}\} \text{ for } i = 1, 2, \dots, 2k$$

where $x_{2k+1} = x_1$. Then

$$\sum_{\ell=1}^{2k} (-1)^{\ell+1} v_i = (x_1 + x_2) - (x_2 + x_3) + \dots - (x_{2k} + x_1) = 0,$$

and the disjoint sets $A = \{v_1, v_3, \dots, v_{2k-1}\}$ and $B = \{v_2, v_4, \dots, v_{2k}\}$ have the same sum. If r is odd, then G is a bipartite graph whose parts have sizes M and N . By the second assertion of Theorem 2.1, $C_{2k} \subseteq G$, and we conclude by the same argument as that presented above. \square

3 An Algorithm for Linear Dependences

From the proof of Theorem 2.2, and the algorithm of Alon-Yuster-Zwick [5, 6] for finding cycles in graphs (or alternatively the proof of the main theorem in [27]), one can give a linear time algorithm for producing a linear dependence of size $2k$ in sets X of size $\Omega\left(n^{\frac{r}{2} + \frac{r}{2k}}\right)$. In fact, the algorithm finds the sets A, B in the statement of Theorem 2.2 in linear time in $|X|$. This is surprisingly independent of the value of k , provided k is a constant as n tends to infinity, and should be compared with the exhaustive search algorithm of all $O(|X|^{2k})$ possible subsets of X of size $2k$. Furthermore, using our proof of Theorem 1.1 for sets of vectors of odd weight, this algorithm can be extended to a $O(|X|^4)$ time algorithm for finding a linear dependence of size $2k$ in a set of vectors of weight r and size $\Omega\left(n^{\frac{r}{2} + \frac{4r}{3k}}\right)$. The details of this algorithm and a discussion of the decision problem for linear dependences are presented in the full version of the paper.

References

- [1] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *J. Algorithms*, 7(4):567–583, 1986.
- [2] N. Alon, O. Goldreich, J. Håstad, and R. Peralta. Simple constructions of almost k -wise independent random variables. *Random Structures Algorithms*, 3(3):289–304, 1992.
- [3] N. Alon, S. Hoory, and N. Linial. The Moore bound for irregular graphs. *Graphs Combin.*, 18(1):53–57, 2002.
- [4] N. Alon and J. H. Spencer. *The probabilistic method*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience [John Wiley & Sons], New York, second edition, 2000. With an appendix on the life and work of Paul Erdős.

- [5] N. Alon, R. Yuster, and U. Zwick. Color-coding. *J.ACM*, 42(4):844–856, 1995.
- [6] N. Alon, R. Yuster, and U. Zwick. Finding and counting given length cycles. *Algorithmica*, 17(3):209–223, 1997.
- [7] L. Bazzi, M. Mahdian, and D. A. Spielman. The minimum distance of Turbo-like codes. Preprint, 2003.
- [8] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon Limit Error Correcting Codes and Decoding: Turbo Codes. In *Proceedings of IEEE International Communications Conference*, pages 1064–1070. 1993.
- [9] C. Bertram-Kretzberg, T. Hofmeister, and H. Lefmann. Sparse 0-1 matrices and forbidden hypergraphs. *Combin. Probab. Comput.*, 8(5):417–427, 1999.
- [10] C. Bertram-Kretzberg and H. Lefmann. MOD_p -tests, almost independence and small probability spaces. *Random Structures Algorithms*, 16(4):293–313, 2000.
- [11] A. Beutelspacher and U. Rosenbaum. *Projective geometry: from foundations to applications*. Cambridge University Press, Cambridge, 1998.
- [12] J. Bondy and M. Simonovits. Cycles of even length in graphs. *J. Combinatorial Theory B*, 16:97–105, 1974.
- [13] M. Breiling. A logarithmic upper bound on the minimum distance of Turbo codes. Preprint, 2001.
- [14] M. C. Davey and D. J. C. MacKay. Low-density parity check codes over $GF(q)$. *IEEE Communications Letters*, 2(6):165–167, 1998.
- [15] P. Erdős, A. Sárközy, and V. T. Sós. On product representations of powers. I. *European J. Combin.*, 16(6):567–588, 1995.
- [16] R. G. Gallager. *Low Density Parity Check Codes*. MIT Press, Cambridge MA, 1963. Research Monograph Series, no. 21.
- [17] S. Hoory. The size of bipartite graphs with a given girth. *J. Combin. Theory Ser. B*, 86(2):215–220, 2002.
- [18] N. Kahale and R. Urbanke. On the minimum distance of parallel and serially concatenated codes. *IEEE Trans. Inform. Theory*. To appear.
- [19] H. Lefmann. Sparse parity-check matrices over finite fields (extended abstract). In *Computing and combinatorics*, volume 2697 of *Lecture Notes in Comput. Sci.*, pages 112–121. Springer, Berlin, 2003.
- [20] H. Lefmann, P. Pudlák, and P. Savický. On sparse parity check matrices. *Des. Codes Cryptogr.*, 12(2):107–130, 1997.

- [21] D. J. C. MacKay. Good error-correcting codes based on very sparse matrices. *IEEE Trans. Inform. Theory*, 45(2):399–431, 1999.
- [22] F. J. MacWilliams and N. J. A. Sloane. *The theory of error-correcting codes. I*. North-Holland Publishing Co., Amsterdam, 1977. North-Holland Mathematical Library, Vol. 16.
- [23] F. J. MacWilliams and N. J. A. Sloane. *The theory of error-correcting codes. II*. North-Holland Publishing Co., Amsterdam, 1977. North-Holland Mathematical Library, Vol. 16.
- [24] A. Naor and J. Verstraëte. A note on bipartite graphs without a $2k$ -cycle. Preprint, 2003.
- [25] M. Sipser and D. A. Spielman. Expander codes. *IEEE Trans. Inform. Theory*, 42(6, part 1):1710–1722, 1996. Codes and complexity.
- [26] D. A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Trans. Inform. Theory*, 42(6, part 1):1723–1731, 1996. Codes and complexity.
- [27] J. Verstraëte. On arithmetic progressions of cycle lengths in graphs. *Combin. Probab. Comput.*, 9(4):369–373, 2000.