Sparse Parity-Check Matrices over Finite Fields (Extended Abstract)

Hanno Lefmann

Fakultät für Informatik, TU Chemnitz, D-09107 Chemnitz, Germany
lefmann@informatik.tu-chemnitz.de

Abstract. For fixed positive integers k, q, r with q a prime power and large m, we investigate matrices with m rows and a maximum number $N_q(m, k, r)$ of columns, such that each column contains at most r nonzero entries from the finite field $GF(q)$ and each k columns are linearly independent over $GF(q)$. For even k we prove the lower bounds $N_q(m, k, r) = \Omega(m^{kr/(2(k-1))})$, and $N_q(m, k, r) = \Omega(m^{(k-1)r/(2(k-2))})$ for odd $k \geq 3$. For $k = 2^i$ and $\gcd(k-1, r) = k-1$ we obtain $N_q(m, k, r) = \Theta(m^{kr/(2(k-1))})$, while for any even $k \geq 4$ and $\gcd(k-1, r) = 1$ we have $N_q(m, k, r) = \Omega(m^{kr/(2(k-1))} \cdot (\log m)^{1/(k-1)})$. For char $(GF(q)) > 2$ we prove that $N_q(m, 4, r) = \Theta(m^{\lceil 4r/3 \rceil/2})$, while for $q = 2^l$ we only have $N_q(m, 4, r) = O(m^{\lceil 4r/3 \rceil/2})$. We can find matrices, fulfilling these lower bounds, in polynomial time. Our results extend and complement earlier results from [4, 14], where the case $q = 2$ was considered.

1 Introduction

For a prime power q, let $GF(q)$ be the finite field with q elements. We consider matrices over $GF(q)$ with k-wise independent columns, i.e. each k columns are linearly independent over $GF(q)$, and each column contains at most r nonzero entries from $GF(q) \setminus \{0\}$. Such matrices are called (k, r)*-matrices*. For a given number m of rows, let $N_q(m, k, r)$ denote the maximum number of columns a (k, r)-matrix can have. Matrices with k-wise independent columns are just parity-check matrices for linear codes with minimum distance at least $k + 1$, hence we investigate sizes of sparse parity-check matrices over $GF(q)$.

Here, k, r, q are always fixed positive integers and m is large. By monotonicity, we have $N_q(m, k + 1, r) \leq N_q(m, k, r)$ for $k = 2, 3, \ldots$. For $q = 2$ it was shown in [14] that $N_2(m, 2k + 1, r) \geq 1/2 \cdot N_2(m, 2k, r)$, hence asymptotically it suffices here to consider even independences. The values of $N_2(m, k, 2)$ are asymptotically equal to the maximum number of edges in a graph on m vertices without any cycles of length at most k. Although graphs without short cycles are widely used, not that much is known on the exact growth rate of $N_2(m, k, 2)$: known is $N_2(m, 4, 2) = \Theta(m^{3/2})$, see [8, 9], as well as $N_2(m, 6, 2) = \Theta(m^{4/3})$ and $N_2(m, 10, 2) = \Theta(m^{6/5})$, see [3, 18]. Improving results by Margulis [17] and Lubotzky, Phillips and Sarnak [15], Lazebnik, Ustimenko and Woldar [13] showed that $N_2(m, 2k, 2) = \Omega(m^{1+2/(3k-3+\varepsilon)})$ with $\varepsilon = 0$ for k odd, and $\varepsilon = 1$ for k even. Concerning upper bounds we have $N_2(m, 2k, 2) = O(m^{1+1/k})$ by the work of Bondy and Simonovits [7]. For $q = 2$ and any fixed $r \geq 1$, $k \geq 4$ even, the following bounds were given by Pudlák, Savický and this author [14]: $N_2(m,k,r) = \Omega(m^{kr/(2(k-1))})$, and, for $k = 2^i$ it is $N_2(m,k,r) = O(m^{\lceil k \cdot r/(k-1) \rceil /2})$. Thus for $k = 2^i$ and $\gcd(k-1,r) = k-1$ lower and upper bounds match. However, for $\gcd(k-1,r) = 1$ the lower bound was improved by Bertram-Kretzberg, Hofmeister and this author [4] to $N_2(m,k,r) = \Omega(m^{kr/(2(k-1))} \cdot (\ln m)^{1/(k-1)})$.

Here we generalize and extend some of these earlier results to arbitrary finite fields $GF(q)$: we obtain the lower bounds $N_q(m,k,r) = \Omega(m^{kr/(2(k-1))})$ for even k, and $N_q(m,k,r) = \Omega(m^{(k-1)r/(2(k-2))})$ for odd $k \geq 3$. For $k = 2^i$ we show that $N_q(m,k,r) = \Theta(m^{kr/(2(k-1))})$ for $\gcd(k-1,r) = k-1$, while for every even $k \geq 4$ with $\gcd(k-1,r) = 1$ we have $N_q(m,k,r) = \Omega(m^{kr/(2(k-1))} \cdot (\log m)^{1/(k-1)})$. Also, for $k = 4$ and char $(GF(q)) > 2$ we prove that $N_q(m,4,r) = \Theta(m^{\lceil 4r/3 \rceil /2})$, while for $q = 2^l$ we can only show that $N_q(m,4,r) = O(m^{\lceil 4r/3 \rceil /2})$. Matrices, which fulfill these lower bounds, can be found in polynomial time. In Section 4 we discuss some applications.

2 Upper Bounds

Since we only care about independencies, *here* in every matrix M all columns are pairwise distinct, in each column the first nonzero entry is 1 and M does not contain an all zeros column, hence $N_q(m,2,r) = \sum_{i=1}^{r} \binom{m}{i} \cdot (q-1)^{i-1} = \Theta(m^r)$.

Lemma 1. *Let M be an $m \times n$-matrix over $GF(q)$ with at most $r \geq 1$ nonzero entries in each column. Then the matrix M contains an $m \times n'$-submatrix M' with the following properties:*

(*i*) *$n' \geq r!/(r^r \cdot q^r) \cdot n$, and*
(*ii*) *there is a partition $\{1,\ldots,m\} = R_1 \cup \ldots \cup R_r$ of the set of row-indices of M' and a sequence $(f_1, f_2, \ldots, f_r) \in (GF(q))^r$ such that for each column in M' for $j = 1,\ldots,r$ the following holds: if $f_j = 0$ then there are only zeros within the rows of R_j, and if $f_j \neq 0$ then there is exactly one entry f_j within the rows of R_j and the other entries in R_j are equal to zero, and*
(*iii*) *the columns of M' are 3-wise independent.*

Proof. Partition at random the set $\{1,\ldots,m\}$ of row-indices of the $m \times n$-matrix M into r parts $R_1,\ldots,R_r$ all of nearly equal size $\lfloor m/r \rfloor$ or $\lceil m/r \rceil$. It is easily seen that the expected number of columns of M with at most one nonzero entry in each row-set R_j, $j = 1,\ldots,r$, is at least $r!/r^r \cdot n$. Take such a subset of columns of M with corresponding partition $\{1,\ldots,m\} = R_1 \cup \ldots \cup R_r$ and call the resulting matrix M^*. For each column in M^* record for $j = 1,\ldots,r$ the occurring nonzero entry f_j within row-set R_j, and set $f_j = 0$, if all entries from R_j are zero. As there are less than q^r such sequences, there are $n' \geq r!/(r^r \cdot q^r) \cdot n$ columns with the same pattern $(f_1,\ldots,f_r)$. These columns form a submatrix M', which fulfills (i) and (ii). Property (iii) easily follows from property (ii). □

Lemma 1 can be made constructive in polynomial time by applying derandomization techniques. From Lemma 1 for fixed $r \geq 1$ and q a prime power we obtain $N_q(m,3,r) = \Theta(m^r)$ and $N_q(m,5,r) = \Theta(N_q(m,4,r))$.

Theorem 1. *Let $k \geq 4$ be even, $r \geq 1$, and q a prime power. Then, for a constant $c > 0$ and $s = 0, \ldots, r-1$, and for integers $m \geq 1$,*

$$N_q(m,k,r) \leq c \cdot \sqrt{m^s \cdot N_q(m,k/2,2r-2s)} + c \cdot m^s. \tag{1}$$

Proof. The proof is similar, but different, to that in [14], where $q = 2$ was considered. Let M be a (k,r)-matrix of dimension $m \times n$, $n = N_q(m,k,r)$, with entries from $GF(q)$. By Lemma 1, the matrix M contains an $m \times n'$-submatrix M', $n' \geq c^* \cdot n$ with $c^* = r!/(r^r \cdot q^r)$, which satisfies (ii), (iii) there.

Put the columns of M' into two matrices M_1 and M_2 of dimensions $m \times n_1$ and $m \times n_2$, respectively, with $n' = n_1 + n_2$. In M_1 put those columns of M' which have with another column from M' at least s nonzero entries at the same positions. In M_2 put the remaining columns, thus $n_2 \leq \sum_{i=1}^{s} \binom{m}{i}$.

Set $[m] := \{1, \ldots, m\}$, let $|c|$ denote the number of nonzero entries in any column c, and for each subset $S \in [[m]]^s$ of row-indices, let $n(S)$ be the number of columns c in M_1 with $c_s \neq 0$ for each position $s \in S$. Using $|c| \geq s$ for each column c in M_1, and then the Cauchy-Schwartz inequality we infer

$$n_1 \leq L := \sum_{S \in [[m]]^s} n(S) = \sum_{c \in M_1} \binom{|c|}{s},$$

$$\sum_{S \in [[m]]^s} \binom{n(S)}{2} \geq \frac{1}{2} \cdot \frac{L \cdot \left(L - \binom{m}{s}\right)}{\binom{m}{s}}. \tag{2}$$

Form a matrix M_1^* by taking all differences $c_i - c_j$, $i < j$, of those columns of M_1, which share at least at s positions the same nonzero entries. Each column in M_1^* contains at most $2r - 2s$ nonzero entries and these columns are $k/2$-wise independent, hence M_1^* has at most $N_q(m,k/2,2r-2s)$ columns. The sum $\sum_{S \in [[m]]^s} \binom{n(S)}{2}$ counts every pair of distinct columns at most $\binom{r-1}{s}$ times, thus

$$\sum_{S \in [[m]]^s} \binom{n(S)}{2} \leq \binom{r-1}{s} \cdot N_q(m,k/2,2r-2s)$$

$$n_1 \leq L \leq \binom{m}{s} + \sqrt{2 \cdot \binom{m}{s} \cdot \binom{r-1}{s} \cdot N_q(m,k/2,2r-2s)} \qquad \text{(by (2))}.$$

With $n_1 + n_2 \geq c^* \cdot n$ and $n_2 \leq \sum_{i=1}^{s} \binom{m}{i}$, the upper bound (1) follows. $\square$

Corollary 1. *Let $k, r \geq 1$ and q a prime power. For positive integers m,*

$$N_q(m,k,2) = O\left(m^{1+2/2^{\lfloor \log_2 k \rfloor}}\right) \tag{3}$$

$$N_q(m,k,r) = O\left(m^{\lceil kr/(k-1) \rceil /2}\right) \qquad \text{if } k = 2^j. \tag{4}$$

Proof. Inequality (3) follows by induction on $\lfloor \log_2 k \rfloor$ from (1) for $s := 1$, and (4) follows by induction on j and (1) for $s := \lfloor \lceil kr/(k-1) \rceil /2 \rfloor$, compare [14]. $\square$

Lemma 2. *Let $k \geq 2$ and q a prime power. For positive integers m,*

$$N_q(m,k,2) \geq (1 - o(1)) \cdot N_2(m,k,2) \; .$$

Proof. This follows by considering the incidence matrix of a graph without cycles of length at most k as a matrix over $GF(q)$. Details are in the full version. $\square$

With Lemma 2 we obtain the following (constructive) lower bounds from graphs, see [3, 13, 18], for every $k \geq 1$ and q a prime power: $N_q(m,4,2) = \Theta(m^{3/2})$, $N_q(m,5,2) = \Theta(m^{3/2})$, $N_q(m,6,2) = \Omega(m^{4/3})$, $N_q(m,10,2) = \Omega(m^{6/5})$, and $N_q(m,2k,2) = \Omega(m^{1+2/(3k-3+\varepsilon)})$ with $\varepsilon \in \{0,1\}$ and $\varepsilon = 1$ iff k is even.

The next lemma shows that asymptotically it suffices to consider $N_q(m,k,r)$ for q a prime. We omit the proof, which uses linear algebra.

Lemma 3. *Let $k \geq 2$, and $l, r \geq 1$, and p a prime. For positive integers m,*

$$N_{p^l}(m,k,r) = \Theta(N_p(m,k,r)) \; .$$

From the results for graphs [3, 7, 18] and by Lemma 3 we obtain the following:

Corollary 2. *Let $q = 2^l$ and $k \geq 1$. For positive integers m,*

$$N_q(m,6,2) = \Theta(m^{4/3}) \qquad \textit{and} \qquad N_q(m,10,2) = \Theta(m^{6/5})$$
$$N_q(m,2k,2) \;=\; O(m^{1+1/k}) \; .$$

Next we consider $(4,r)$-matrices over $GF(q)$.

Lemma 4. *Let $GF(q)$ be a finite field with char $(GF(q)) > 2$. Let M' be an $m \times n$-matrix over $GF(q)$ with exactly r nonzero entries in each column, such that Lemma 1 (ii), (iii) are satisfied. Let $F_1', \ldots, F_n'$ be the sets of positions of the nonzero entries in the n columns of M'. If for no four sets both $F_g' \cup F_h' = F_i' \cup F_j'$ and $F_g' \cap F_h' = F_i' \cap F_j'$ are fulfilled, then the columns of M' are 4-wise independent.*

Proof. Assume that some columns $a_1, \ldots, a_4$ of M', with corresponding sets $F_1', \ldots, F_4'$, are linearly dependent over $GF(q)$, i.e. $\sum_{i=1}^{4} \lambda_i \cdot a_i = 0$ for some $\lambda_1, \ldots, \lambda_4 \in GF(q) \setminus \{0\}$. Let $F_i := F_i' \setminus (F_1' \cap \ldots \cap F_4')$ for $i = 1, \ldots, 4$.

Lemma 5. *For $1 \leq i < j < k \leq 4$ it is $F_i \cap F_j \cap F_k = \emptyset$.*

Proof. This follows with Lemma 1 (ii). Details are in the full version. $\square$

Assume w.l.o.g. that $F_1 \cap F_2 \neq \emptyset$. By Lemma 5 we have $F_3 \not\subseteq F_1 \cup F_2$, hence $F := F_3 \setminus (F_1 \cup F_2) \neq \emptyset$. The dependence of $a_1, \ldots, a_4$ implies $F_4 \setminus (F_1 \cup F_2) = F$, thus $\lambda_3 = -\lambda_4$. If $F_3 \cap (F_1 \setminus F_2) \neq \emptyset$ and $F_4 \cap (F_1 \setminus F_2) \neq \emptyset$, then $\lambda_3 = -\lambda_1$ and $\lambda_4 = -\lambda_1$, which shows $\lambda_3 = \lambda_4 = 0$ for char $(GF(q)) > 2$, a contradiction. Hence, $F_3 \cap (F_1 \setminus F_2) = \emptyset$ or $F_4 \cap (F_1 \setminus F_2) = \emptyset$. Then w.l.o.g. $F_3 = F \cup (F_2 \setminus F_1)$ and $F_4 = F \cup (F_1 \setminus F_2)$, hence $F_1 \cup F_3 = F_2 \cup F_4$ and $F_1 \cap F_3 = F_2 \cap F_4$, which contradicts the assumption. $\square$

Frankl and Füredi [11] constructed a family $\mathcal{F} \subseteq [[m]]^r$ containing no four sets $F_1, \ldots, F_4 \in \mathcal{F}$ with $F_1 \cup F_2 = F_3 \cup F_4$ and $F_1 \cap F_2 = F_3 \cap F_4$, where $|\mathcal{F}| = \Theta(m^{\lceil 4r/3 \rceil/2})$, as follows. Let $r = 3t + 1$ (For other values of ($r \bmod 3$) the construction is similar.), and let K be any field with $m/2 \leq |K| \leq m$. For a subset $X = \{x_1, \ldots, x_g\} \subseteq K$ let $s_i(X) := \sum_{I \in [[g]]^i} \prod_{j \in I} x_j$. For $h \geq 1$ define an $h \times h$-matrix $M_h(X)$ with entries $m_{i,j} = s_{2i-j}(X)$ (where $s_l(X) = 0$ for $l < 0$ or $l > |X|$). Using an averaging argument, for suitable $c_2, c_4, \ldots, c_{2t} \in K$ the family $\mathcal{F} \subseteq [K]^r$ is defined as follows: $X = \{x_1, \ldots, x_r\} \in \mathcal{F}$ if $s_{2i}(X) = c_{2i}$ for $i = 1, \ldots, t$ and $\det(M_h(S)) \neq 0$ for every subset $S \subseteq X$ and $h = 1, \ldots, |S| - 1$. This yields a polynomial time (semi-) construction.

Theorem 2. *Let $r \geq 1$ and q a prime power. For positive integers m,*

$$N_q(m, 4, r) = \begin{cases} \Theta\left(m^{\lceil 4r/3 \rceil/2}\right) & \text{if char } (GF(q)) > 2 \\ O\left(m^{\lceil 4r/3 \rceil/2}\right) & \text{if char } (GF(q)) = 2. \end{cases}$$

For $q = 2$ this was shown in [14]. For $q = 2^l$ and $r \equiv 0 \bmod 3$ lower and upper bound match by Theorem 5, compare also Theorem 3.

Proof. For the upper bounds see (4). For the lower bound, let $\mathcal{F} = \{F_1, \ldots, F_n\} \subseteq [[m]]^r$ with $n = \Theta(m^{\lceil 4r/3 \rceil/2})$, such that for no four sets $F_1, \ldots, F_4 \in \mathcal{F}$ it is $F_1 \cup F_2 = F_3 \cup F_4$ and $F_1 \cap F_2 = F_3 \cap F_4$. Such a family exists by the results in [11]. Define an $m \times n$-matrix M with columns $c_1, \ldots, c_n \in \{0, 1\}^m$: in column c_i put a 1 at position s iff $s \in F_i$, $i = 1, \ldots, n$. By Lemma 1 we obtain an $m \times n'$-submatrix M' of M such that (i) – (iii) there are satisfied. By Lemma 4, the columns of M' are 4-wise independent and the lower bound follows. □

3 Lower Bounds

For proving lower bounds on $N_q(m, k, r)$, we will look for a large independent set in a suitable hypergraph. A *hypergraph* $\mathcal{G} = (V, \mathcal{E})$ has vertex set V and edge set $\mathcal{E}$ where $E \subseteq V$ for every edge $E \in \mathcal{E}$. A hypergraph $\mathcal{G} = (V, \mathcal{E})$ is *l-uniform*, if the edge set $\mathcal{E}$ contains only l-element edges, i.e. $\mathcal{E} \subseteq [V]^l$. An *independent set* in $\mathcal{G} = (V, \mathcal{E})$ is a subset $I \subseteq V$ which contains no edges from $\mathcal{E}$.

Theorem 3. *Let $k \geq 4$, $r \geq 2$, and q a prime power. For positive integers m,*

$$N_q(m, k, r) = \Omega\left(m^{\frac{kr}{2(k-1)}} \cdot (\log m)^{\frac{1}{k-1}}\right) \quad \text{for } k \text{ even, } \gcd(k-1, r) = 1 \quad (5)$$

$$N_q(m, k, r) = \Omega\left(m^{\frac{(k-1)r}{2(k-2)}} \cdot (\log m)^{\frac{1}{k-2}}\right) \quad \text{for } k \text{ odd, } \gcd(k-2, r) = 1. \quad (6)$$

For $q = 2$ this was shown in [4], thus (5), (6) hold for $q = 2^l$ by Lemma 3.

Proof. Partition the set $\{1, \ldots, m\}$ of row-indices into r subsets $R_1, \ldots, R_r$ of sizes $\lfloor m/r \rfloor$ or $\lceil m/r \rceil$, and fix a sequence $(f_1 = 1, f_2, \ldots, f_r) \in (GF(q) \setminus \{0\})^r$ of nonzero elements. Let $C_q(m, r)$ be the set of all column vectors of length m

which contain within each row-set R_j exactly one nonzero entry $f_j \in GF(q)\setminus\{0\}$, $j = 1, \ldots, r$. Hence $|C_q(m,r)| = c \cdot m^r$ for some constant $c > 0$.

Form a hypergraph $\mathcal{G} = (V, \mathcal{E}_3 \cup \ldots \cup \mathcal{E}_k)$ with vertex set $V = C_q(m,r)$, where an i-element subset $\{a_1, \ldots, a_i\}$ of V, $i = 3, \ldots, k$, is an edge in this hypergraph $\mathcal{G}$, that is $\{a_1, \ldots, a_i\} \in \mathcal{E}_i$, if and only if $a_1, \ldots, a_i$ are linearly dependent over $GF(q)$ but any $(i-1)$ columns of $a_1, \ldots, a_i$ are linearly independent.

We will prove a lower bound on the size of a maximum independent set in $\mathcal{G}$, which yields a set of k-wise independent columns. First we will bound from above the numbers $|\mathcal{E}_i|$, $i = 3, \ldots, k$, of i-element edges in $\mathcal{G}$. For a subset $E \in [C_q(m,r)]^i$ of i column vectors, the $m \times i$-matrix $M(E)$ formed by these columns has exactly $i \cdot r$ nonzero entries. If $E \in \mathcal{E}_i$, then in $M(E)$ each row-set R_j contains at most $\lfloor i/2 \rfloor$ non-zero rows, hence $M(E)$ contains at most $\lfloor i/2 \rfloor \cdot r$ nonzero rows. Thus, for some constants $c_i > 0$, $i = 3, \ldots, k$, we have

$$|\mathcal{E}_i| \leq \binom{\lceil m/r \rceil}{\lfloor i/2 \rfloor}^r \cdot \binom{i \cdot \lfloor i/2 \rfloor \cdot r}{ir} \leq c_i \cdot m^{\lfloor i/2 \rfloor \cdot r}. \tag{7}$$

Next we consider only edges in $\mathcal{E}_l$, where $l := k$ for k even and $l := k-1$ for k odd, hence $l \geq 4$ is *always even*. For a subset $J \in [C_q(m,r)]^j$, $j = 2, \ldots, l-1$, let $p(J)$ be the number of nonzero rows in the matrix $M(J)$, and let $p_1(J)$ be the number of rows in $M(J)$ with exactly one nonzero entry. Let $b(J)$ be the number of subsets $S \in [C_q(m,r)]^{l-j}$ where $J \cup S \in \mathcal{E}_l$. For $J \cup S \in \mathcal{E}_l$ in each row in $M(J)$ with exactly one nonzero entry there must be at least one such nonzero entry in the same row in $M(S)$. Let $M(J)$ have the nonzero rows $1, \ldots, p(J)$, say. Each nonzero row $s > p(J)$ in $M(S)$ has at least two nonzero entries, as the columns in $J \cup S$ are linearly dependent, hence there are at most $\lfloor ((l-j)r - p_1(J))/2 \rfloor$ such rows in $M(S)$, i.e. at most $\binom{m-p(J)}{\lfloor ((l-j)r - p_1(J))/2 \rfloor}$ choices for these, thus for a constant $c_p > 0$:

$$b(J) \leq c_p \cdot m^{\lfloor \frac{(l-j)r - p_1(J)}{2} \rfloor}. \tag{8}$$

A 2-*cycle* in an l-uniform hypergraph $\mathcal{G} = (V, \mathcal{E})$ is a pair $\{E, E'\}$ of distinct edges $E, E' \in \mathcal{E}$ with $|E \cap E'| \geq 2$. We will apply a result of Ajtai, Komlós, Pintz, Spencer and Szemerédi [1], originally an existence result, in the sequel extended and turned into a deterministic polynomial time algorithm in [10, 5].

Theorem 4. *Let $l \geq 3$ be a fixed integer. Let $\mathcal{G} = (V, \mathcal{E})$ be an l-uniform hypergraph on $|V| = N$ vertices with average degree $t^{l-1} := l \cdot |\mathcal{E}|/|V|$. If the hypergraph $\mathcal{G} = (V, \mathcal{E})$ contains no 2-cycles, then one can find for any fixed $\delta > 0$ in $\mathcal{G}$ in time $O(N \cdot t^{l-1} + N^3/t^{3-\delta})$ an independent set of size $\Omega(N/t \cdot (\log t)^{1/(l-1)})$. The assertion also holds, if t^{l-1} is only an upper bound on the average degree.*

For $j = 2, \ldots, l-1$ and $u = 0, \ldots, jr$, let $s_{2,j}(u; \mathcal{G}_l)$ be the number of 2-cycles $\{E, E'\}$ in $\mathcal{G}_l = (V, \mathcal{E}_l)$ with $|E \cap E'| = j$ and $p_1(E \cap E') = u$. The numbers $p_{j,u}(V)$ of subsets $J \in [C_q(m,r)]^j$ with $p_1(J) = u$ satisfy for constants $c^*_{j,u}, c_{j,u} > 0$:

$$p_{j,u}(V) \leq c^*_{j,u} \cdot \binom{m}{u} \cdot \binom{m-u}{\lfloor (jr-u)/2 \rfloor} \leq c_{j,u} \cdot m^{u + \lfloor \frac{jr-u}{2} \rfloor}, \tag{9}$$

since the matrix $M(J)$ has u rows with exactly one nonzero entry and the remaining $jr-u$ nonzero entries are contained in rows with at least two nonzero entries, thus by (8) and (9) for some constant $C_1 > 0$:

$$s_{2,j}(u;\mathcal{G}_l) \le \sum_{J \in [C_q(m,r)]^j;\, p_1(J)=u} \binom{b(J)}{2} \le C_1 \cdot m^{2\cdot \lfloor \frac{(l-j)r-u}{2} \rfloor + u + \lfloor \frac{jr-u}{2} \rfloor}. \quad (10)$$

By (7) the *average degree* t^{l-1} of the l-uniform hypergraph $\mathcal{G}_l = (V, \mathcal{E}_l)$ satisfies $t^{l-1} = (l \cdot |\mathcal{E}_l|)/|V| \le (l \cdot c_l \cdot m^{\lfloor l/2 \rfloor \cdot r})/(c \cdot m^r)$, hence $t \le t_0 := C_2 \cdot m^{(\lfloor l/2 \rfloor \cdot r - r)/(l-1)}$ for a constant $C_2 > 0$. To apply Theorem 4, we choose a random subset $V^* \subseteq V$ by picking vertices at random from V, independently of each other with probability $p := t_0^{-1} \cdot m^{\varepsilon}$ for some constant $\varepsilon > 0$, heading for an l-uniform hypergraph without any 2-cycles. We estimate the expected values $E(\cdot)$ of some parameters of the induced random hypergraph $\mathcal{G}^* = (V^*, \mathcal{E}_3^* \cup \ldots \cup \mathcal{E}_k^*)$ with $\mathcal{E}_i^* := \mathcal{E}_i \cap [V^*]^i$, $i = 3, \ldots, k$, using (7) for some constants $c^*, c_i^* > 0$ as follows:

$$E(|V^*|) = p \cdot |C_q(m,r)| \ge c^* \cdot m^{r - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} + \varepsilon} \quad (11)$$

$$E(|\mathcal{E}_i^*|) = p^i \cdot |\mathcal{E}_i| \le p^i \cdot c_i \cdot m^{\lfloor \frac{i}{2} \rfloor \cdot r} \le c_i^* \cdot m^{\lfloor \frac{i}{2} \rfloor \cdot r - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot i + i \cdot \varepsilon}. \quad (12)$$

With (9), (10) we infer for $j = 2, \ldots, l-1$ and $u = 0, \ldots, j \cdot r$, where $\mathcal{G}_l^* = (V^*, \mathcal{E}_l^*)$, for some constants $C_0^*, C_1^* > 0$:

$$E(p_{j,u}(V^*)) = p^j \cdot p_{j,u}(V) \le C_0^* \cdot m^{u + \lfloor \frac{jr-u}{2} \rfloor - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot j + j \cdot \varepsilon} \quad (13)$$

$$\begin{aligned} E(s_{2,j}(u;\mathcal{G}_l^*)) &= p^{2l-j} \cdot s_{2,j}(u;\mathcal{G}_l) \le \\ &\le C_1^* \cdot m^{2 \cdot \lfloor \frac{(l-j)r-u}{2} \rfloor + u + \lfloor \frac{jr-u}{2} \rfloor - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot (2l-j) + (2l-j) \cdot \varepsilon}. \quad (14) \end{aligned}$$

By (11)–(14), Markov's and Chebychev's inequality, there exists a subhypergraph $\mathcal{G}^* = (V^*, \mathcal{E}_3^* \cup \ldots \cup \mathcal{E}_k^*)$ of $\mathcal{G}$ which satisfies for $i = 2, \ldots, k$, and $j = 2, \ldots, l-1$, and $u = 0, \ldots, jr$, (using simply the same notation for the constant factors)

$$|V^*| \ge c^* \cdot m^{r - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} + \varepsilon} \quad (15)$$

$$|\mathcal{E}_i^*| \le c_i^* \cdot m^{\lfloor \frac{i}{2} \rfloor \cdot r - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot i + i \cdot \varepsilon} \quad (16)$$

$$s_{2,j}(u;\mathcal{G}_l^*) \le C_1^* \cdot m^{2 \cdot \lfloor \frac{(l-j)r-u}{2} \rfloor + u + \lfloor \frac{jr-u}{2} \rfloor - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot (2l-j) + (2l-j) \cdot \varepsilon} \quad (17)$$

$$p_{j,u}(V^*) \le C_0^* \cdot m^{u + \lfloor \frac{jr-u}{2} \rfloor - \frac{\lfloor l/2 \rfloor \cdot r - r}{l-1} \cdot j + j \cdot \varepsilon}. \quad (18)$$

Lemma 6. *For $k \ge 4$ and $\varepsilon = r/(2k^2)$ it holds:*

$$|\mathcal{E}_i^*| = o(|V^*|) \text{ for every } i \ne l. \quad (19)$$

Proof. By (15) and (16), since l is even, we have $|E_i^*| = o(|V^*|)$, $i \ne l$, if

$$r - \lfloor i/2 \rfloor \cdot r + (i-1)(l-2)r/(2(l-1)) - (i-1) \cdot \varepsilon > 0\,. \quad (20)$$

Inequality (20) holds, if $(l-i)r/(2(l-1)) - (i-1) \cdot \varepsilon > 0$, which is fulfilled for $i = 2, \ldots, l-1$ and $0 < \varepsilon < r/(2l^2)$. For $i > l$, which is only possible for $i = k$ odd and $l = k-1$, inequality (20) is equivalent to $(k-3)r/(2(k-2)) - (k-1) \cdot \varepsilon > 0$, which with $k \ge 4$ holds for $0 < \varepsilon \le r/(2k^2)$, thus (19) is satisfied. $\square$

We have the following consequence:

Theorem 5. *Let $k \geq 4$, $r \geq 1$, and q a prime power. For positive integers m,*

$$N_q(m,k,r) = \begin{cases} \Omega\left(m^{\frac{kr}{2(k-1)}}\right) & \textit{if } k \textit{ is even,} \\ \Omega\left(m^{\frac{(k-1)r}{2(k-2)}}\right) & \textit{if } k \textit{ is odd.} \end{cases}$$

Thus, for $k = 2^i$ and $\gcd(k-1, r) = k-1$ lower and upper bound (4) match, i.e. $N_q(m,k,r) = \Theta(m^{kr/(2(k-1))})$. Matrices satisfying the bounds in Theorem 5 can be found in polynomial time by using the method of conditional probabilities.

Proof. For every $i \neq l$ and $\varepsilon = r/(2k^2)$ we have $|\mathcal{E}_i^*| = o(|V^*|)$ by (19), and we remove one vertex from each edge $E \in \mathcal{E}_i^*$, and we obtain a subset $V^{**} \subseteq V^*$ with $|V^{**}| \geq (c^*/2) \cdot m^{lr/(2(l-1))+\varepsilon}$, such that the subhypergraph $\mathcal{G}^{**} = (V^{**}, [V^{**}]^l \cap \mathcal{E}_l^*)$ of $\mathcal{G}^*$ is l-uniform with $|[V^{**}]^l \cap \mathcal{E}_l^*| \leq c_l^* \cdot m^{lr/(2(l-1))+l\cdot\varepsilon}$. Again we choose a random subset $V^{***} \subseteq V^{**}$ by picking vertices from V^{**} uniformly at random, independently of each other with probability $p := c_h \cdot m^{-\varepsilon}$, $c_h := (c^*/(4c_l^*))^{1/(l-1)}$. We estimate $E(|V^{***}|) - E(|[V^{***}]^l \cap \mathcal{E}_l^*|)$, hence there exists a subset $V^{***} \subseteq V^{**}$ such that

$$|V^{***}| - |[V^{***}]^l \cap \mathcal{E}_l^*| \geq (c_h \cdot c^*/4) \cdot m^{lr/(2(l-1))}.$$

We delete one vertex from every edge in $[V^{***}]^l \cap \mathcal{E}_l^*$ and we obtain an independent set I in $\mathcal{G}$ with $|I| = \Omega(m^{lr/(2(l-1))})$, and the lower bounds follow by inserting $l := k$ for k even, and $l := k-1$ for k odd.

We remark, that we could have derived the bounds in Theorem 5 already from (7), by picking vertices right away from the set V at random, independently from each other, each with probability $p := c_g \cdot t_0^{-1}$ for some constant $c_g > 0$. □

Lemma 7. *For $\varepsilon = 1/(4k^2)$, $j = 2, \ldots, l-1$, and $\gcd(l-1, r) = 1$ it is*

$$min_{\ u=0,\ldots,jr}\ \{p_{j,u}(V^*), s_{2,j}(u;\mathcal{G}_l^*)\} = o(|V^*|). \tag{21}$$

Proof. By (15) and (17), we have $s_{2,j}(u;\mathcal{G}_l^*) = o(|V^*|)$, $j = 2, \ldots, l-1$, if

$$0 > (l-1)\cdot r - jr/2 - u/2 - (l-2)(2l-j-1)r/(2(l-1)) + (2l-j-1)\cdot\varepsilon$$

which holds for $u > (l-j)r/(l-1) + 2\cdot(2l-j-1)\cdot\varepsilon$.

For l even by (15) and (18) we have $p_{j,u}(V^*) = o(|V^*|)$, $j = 2, \ldots, l-1$, if

$$u/2 + jr/2 - r - (l-2)(j-1)r/(2(l-1)) + (j-1)\cdot\varepsilon < 0$$

which holds for $u < (l-j)r/(l-1) - 2\cdot(j-1)\cdot\varepsilon$.

For $j = 2, \ldots, l-1$ and $\gcd(l-1, r) = 1$, the quotients $(l-j)r/(l-1)$ are never integers, in particular, they are at least $1/(l-1)$ apart from the nearest integer. For $\varepsilon := 1/(4k^2)$, the statement *'for $u > (l-j)r/(l-1) + 2\cdot(2l-j-1)\cdot\varepsilon$ or $u < (l-j)r/(l-1) - 2\cdot(j-1)\cdot\varepsilon$'* reads as *'for $u = 0, \ldots, jr$'*, and (21) follows. □

Let $\varepsilon := 1/(4k^2)$ and $u_0(j) := \lfloor (l-j)r/(l-1) \rfloor$. For $i \neq l$ we delete one vertex from each edge $E \in \mathcal{E}_i^*$ and, by Lemma 6, we obtain a subset $V^{**} \subseteq V^*$ with $|V^{**}| = (1-o(1)) \cdot |V^*|$. The resulting induced subhypergraph of $\mathcal{G}^{**}$ on V^{**} is l-uniform. For $j = 2, \ldots, l-1$ and $u > u_0(j)$ we delete one vertex from each 2-cycle $\{E, E'\} \in [\mathcal{E}_l^* \cap [V^{**}]^l]^2$ with $|E \cap E'| = j$ and $p_1(E \cap E') = u$. For $u \leq u_0(j)$ we remove one vertex from each subset $J \in [V^{**}]^j$ with $p_1(J) = u$. By Lemma 7 we obtain a subset $V^{***} \subseteq V^{**}$, which does not contain any 2-cycles and satisfies $|V^{***}| = (1-o(1)) \cdot |V^*|$. We apply Theorem 4 to our hypergraph $\mathcal{G}^{***} = (V^{***}, [V^{***}]^l \cap \mathcal{E}_l^*)$ with average degree $t_*^{l-1} \leq c_0 \cdot (1+o(1)) \cdot (p \cdot t_0)^{l-1}$ for some constant $c_0 > 0$, and we obtain an independent set $I \subseteq V^{***}$ of size

$$|I| = \Omega\left(|V^{***}|/(p \cdot t_0) \cdot (\log(p \cdot t_0))^{\frac{1}{l-1}}\right) = \Omega\left(m^{\frac{lr}{2(l-1)}} \cdot (\log m)^{\frac{1}{l-1}}\right) ,$$

which yields the desired lower bounds (5), (6) by inserting the appropriate value of l, i.e. $l := k$ for k even, and $l := k-1$ for k odd.

With the method of conditional probabilities, the running time is essentially bounded by the number $s_{2,2}(u_0(2)+1; \mathcal{G}_l) = O(m^{(l-3)r + r/(l-1)})$ of 2-cycles. Compared with $N^3/t_0^{3-3\delta}$ in Theorem 4, with $t_0 = \Theta(m^\varepsilon)$ (else, for $t_0 = o(m^\varepsilon)$, we can improve (5), (6)), we get the time bound $O(m^{(l-3)r + r/(l-1)})$ for $l \geq 6$.

Remark: All calculations are valid, if we pick the columns uniformly at random according to a $(2l-2)$-wise independent distribution. For simulating this, there exist sample spaces of sizes $O(m^{r(4l-4)})$, see [12], hence with this we also obtain polynomial running time. This finishes the proof of Theorem 3. □

4 Concluding Remarks

Some of the following possible applications have been stated already in [14] for the case $q = 2$ and can be extended to arbitrary prime powers q. For example, we can extend the length of a linear code, but we reduce its minimum distance:

Proposition 1. *Let A be an $l \times m$ parity-check matrix over $GF(q)$ of a linear code of length m with minimum distance at least $kr+1$, and let B be a (k,r)-matrix of dimension $m \times n$ over $GF(q)$. Then the matrix-product $A \times B$ is a parity-check matrix of a code of length n with minimum distance at least $k+1$.*

Proposition 2. *Let A be an $l \times m$-matrix over $GF(q)$ with kr-wise independent columns, and let B be a (k,r)-matrix over $GF(q)$ with dimension $m \times n$. Then the columns in the matrix $A \times B$ are k-wise independent.*

We can use sparse matrices to construct small probability spaces, see [2, 6]:

Proposition 3. *Let M be a (k,r)-matrix over $GF(q)$ of dimension $m \times n$.*

(i) If $X = (X_1, \ldots, X_m)$ is a kr-wise ε-biased random vector over $GF(q)$, then the vector $Y = (Y_1, \ldots, Y_n) = X \times M$ is k-wise ε-biased.
(ii) If $S \subseteq (GF(q))^m$ is a kr-wise ε-biased sample space, then the sample space $T = \{s \times M \mid s^T \in S\} \subseteq (GF(q))^n$ is k-wise ε-biased, hence also $(2 \cdot \varepsilon \cdot (1 - q^{-k})/q, k)$-independent.

Remarks: We can construct large (k, r)-matrices in polynomial time, but our time bound increases (in the exponent) with k, r. For possibly practical applications (quite often algorithms run fast on such matrices) and further studies explicit constructions of (k, r)-matrices are desirable. The constructions in [14] for $q = 2$ can be adapted to arbitrary finite fields, however they yield weaker lower bounds on $N_q(m, k, r)$ than we proved. Apart from $(4, 2)$-matrices over $GF(q)$, the only known asymptotically optimal construction is for $(4, r)$-matrices. Also one might like to investigate possible connections to low-density codes [16].

References

1. M. Ajtai, J. Kómlos, J. Pintz, J. Spencer and E. Szemerédi, Extremal uncrowded hypergraphs, J. Comb. Theory A 32, 1982, 321-335.
2. N. Alon, O. Goldreich, J. Håstad and R. Peralta, Simple constructions of almost k-wise independent random variables, Rand. Struct. & Algorithms 3, 1992, 289-304, and 4, 1993, 119-120.
3. C. T. Benson, Minimal regular graphs of girth eight and twelve, Canad. J. Mathematics 18, 1966, 1091-1094.
4. C. Bertram-Kretzberg, T. Hofmeister and H. Lefmann, Sparse 0-1-matrices and forbidden hypergraphs, Comb., Prob. and Computing 8, 1999, 417-427.
5. C. Bertram-Kretzberg and H. Lefmann, The algorithmic aspects of uncrowded hypergraphs, SIAM J. Computing 29, 1999, 201-230.
6. C. Bertram-Kretzberg and H. Lefmann, MOD_p-tests, almost independence and small probability spaces, Rand. Struct. & Algorithms 16, 2000, 293-313.
7. A. Bondy and M. Simonovits, Cycles of even length in graphs, J. Comb. Theory Ser. B 16, 1974, 97-105.
8. W. G. Brown, On graphs that do not contain a Thomsen graph, Canad. Math. Bulletin 9, 1966, 281-289.
9. P. Erdös, A. Rényi and V. T. Sós, On a problem of graph theory, Stud. Sci. Math. Hungarica 1, 1966, 213-235.
10. A. Fundia, Derandomizing Chebychev's inequality to find independent sets in uncrowded hypergraphs, Rand. Struct. & Algorithms 8, 1996, 131-147.
11. P. Frankl and Z. Füredi, Union-free families of sets and equations over fields, J. Numb. Theory 23, 1986, 210-218.
12. H. Karloff and Y. Mansour, On construction of k-wise independent random variables, Proc. 26th STOC, 1994, 564-573.
13. F. Lazebnik, V. A. Ustimenko and A. J. Woldar, A new series of dense graphs of high girth, Bull. (New Series) of the AMS 32, 1995, 73-79.
14. H. Lefmann, P. Pudlák and P. Savický, On sparse parity-check matrices, Designs, Codes and Cryptography 12, 1997, 107-130.
15. A. Lubotzky, R. Phillips and P. Sarnak, Ramanujan graphs, Combinatorica 8, 1988, 261-277.
16. M. Luby, M. Mitzenmacher, A. Shokrollahi and D. Spielman, Analysis of low-density codes and improved designs using irregular graphs, Proc. 30th STOC, 1998, 249-258.
17. G. A. Margulis, Explicit group theoretical construction of combinatorial schemes and their application to the design of expanders and concentrators, J. Probl. Inform. Transm. 24, 1988, 39-46.
18. R. Wenger, Extremal graphs with no C_4's, C_6's or C_{10}'s, J. Comb. Theory Ser. B 52, 1991, 113-116.