Combinatorics, Probability and Computing (2000) 9, 381–390. Printed in the United Kingdom

String Quartets in Binary

NOGA ALON,[1]† JÁNOS KÖRNER[2] and ANGELO MONTI[2]

[1] Department of Mathematics, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel
(e-mail: noga@math.tau.ac.il)

[2] Department of Computer Science, University of Rome I 'La Sapienza', Via Salaria 113, 00198 Roma, Italy
(e-mail: korner@dsi.uniroma1.it, monti@dsi.uniroma1.it)

Received 10 June 1999; revised 27 October 1999

Let $M(n, A)$ denote the maximum possible cardinality of a family of binary strings of length n, such that for every four distinct members of the family there is a coordinate in which exactly two of them have a 1. We prove that $M(n, A) \leqslant 2^{0.78n}$ for all sufficiently large n. Let $M(n, C)$ denote the maximum possible cardinality of a family of binary strings of length n, such that for every four distinct members of the family there is a coordinate in which exactly one of them has a 1. We show that there is an absolute constant $c < 1/2$ such that $M(n, C) \leqslant 2^{cn}$ for all sufficiently large n. Some related questions are discussed as well.

1. Introduction

Many problems in extremal set theory can be formulated in terms of finding, for fixed k and n, the maximum number of binary strings of length n with the property that no k of them should form a specific 'forbidden' configuration. Sperner's theorem [17] is the answer in the most fundamental and elementary case (with $k = 2$). For $k = 3$ we already have a wealth of intriguing, well-known and unsolved problems of this kind, most of which have been studied extensively in different and often applied contexts. These include strong Δ-systems [4], cancellative families [6], superimposed codes [5] (*cf.* also [3]) and qualitatively 3-independent bipartitions [7]. Until a few months ago all of these problems had one thing in common: not even the exponential growth rate of the maximum number of n-strings with the required property was known. The breakthrough occurred with cancellative set families when Shearer [15] disproved the corresponding conjecture of Erdős and Katona [6] and this led the way to Tolhuizen's beautiful discovery [18] that the Frankl–Füredi upper bound [6] is tight. The construction in [18] based on cosets of randomly chosen

† Research supported in part by a USA–Israeli BSF grant and by a grant from the Israel Science Foundation.

linear codes gives us the first precise asymptotics for a nontrivial problem in this context. Some of the above authors, in particular Dyachkov and Rykov [3] and Tolhuizen [18], use the language of the Shannon Theory of Information; for information-theoretic aspects and the double life of such problems see the survey [10].

Given the difficulty and the unresolved status of the problems for $k = 3$ it might seem odd to look into problems about forbidden configurations of 4 strings. However, as we shall see, the parity of k does play a role here and things change for this new case. We restrict our attention to the following setting. Let $A \subseteq \{0,1\}^4$ be arbitrary and let $M(n, A)$ be the maximum cardinality of a set $\mathscr{C} \subseteq \{0,1\}^n$ with the property that, for every ordered fourtuple $(\mathbf{w}, \mathbf{x}, \mathbf{y}, \mathbf{z})$ of them, there is at least one coordinate i such that $(w_i, x_i, y_i, z_i) \subseteq A$.

If A is the set of all the 6 fourtuples with equal numbers of zeros and ones, then $M(n, A)$ can be reinterpreted as the maximum number of binary n-strings with the property that, for every four of them, there is a coordinate representing a bipartition halving the fourtuple.

Theorem 1.1.

$$\frac{1}{3}\log\frac{8}{5} \leqslant \limsup_{n\to\infty} \frac{1}{n}\log M(n, A) \leqslant 0.773\ldots.$$

Here and henceforth, all logarithms and exponentials are to the base 2.

Similar results for the case when every fourtuple has to be halved in at least two (three) different ways were obtained by Simonyi and Körner [11]. In reaction to that paper Vera T. Sós asked in 1987 whether the above lim sup is strictly less than 1. Our first result gives a positive answer to her question. If, on the other hand, B is the set of all the 8 fourtuples with an odd number of ones, then $M(n, B)$ can be reinterpreted as the maximum number of binary n-strings with the property that any two of them have a different sum modulo 2. For this case Lindström [13] proved the following.

Theorem L. ([13])

$$\lim_{n\to\infty} \frac{1}{n}\log M(n, B) = \frac{1}{2}.$$

The construction part of his result is based on what is called 2-error correcting BCH codes in algebraic coding theory; once again, information theory is lingering around. In view of the last theorem it is interesting to know that, if C is the set of all the 4 fourtuples with a single 1, then we can only construct considerably fewer sequences, and in particular, we shall show the following.

Theorem 1.2.

$$\frac{1}{3}(6 - \log 37) \leqslant \limsup_{n\to\infty} \frac{1}{n}\log M(n, C) < \frac{1}{2}.$$

The proof of Theorem 1.1 relies on Sauer's lemma [14], also known as the Shelah–Perles theorem [16] and the Vapnik–Chervonenkis lemma [19] (all of whom proved it independently and about the same time), while the proof of Theorem 1.2 uses the

bounding technique based on the sub-additivity of graph entropy, introduced in [9]. As is often the case in these problems, the lower bounds are obtained via simple probabilistic constructions.

2. Proof of Theorem 1.1

The lower bound follows by a routine random choice argument, using the so-called alteration method (see, *e.g.*, [1, Chapter 3]). Here are the details. Let F be a random family of N binary strings of length n, where each member of F is chosen, randomly, independently and uniformly, among all binary strings of length n. The expected number of fourtuples in F for which there is no coordinate in which exactly two of them have a 1 is $\binom{N}{4}(10/16)^n$. The expected number of pairs of identical strings in F is $\binom{N}{2}(1/2)^n$. By linearity of expectation, there is a choice of N strings in which the number of such forbidden fourtuples plus the number of such forbidden pairs is at most $\binom{N}{4}(10/16)^n + \binom{N}{2}(1/2)^n$, and by deleting an arbitrarily chosen string from each such forbidden configuration we conclude that, for every N,

$$M(n, A) \geqslant N - \binom{N}{4}(10/16)^n - \binom{N}{2}(1/2)^n.$$

Substituting $N = \lfloor (16/10)^{n/3} \rfloor$, it follows that $M(n, A) \geqslant c(16/10)^{n/3}$ for some absolute positive constant c, supplying the desired lower bound.

We next prove the upper bound. Let $\mathscr{C}_n \subseteq \{0,1\}^n$ achieve

$$\mathscr{C}_n = M(n, A),$$

and write

$$c = \limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n|.$$

By Sauer's lemma, there is a set of coordinates $D_n \subseteq [n]$ satisfying

$$\lim_{n\to\infty} \frac{|D_n|}{n} \geqslant h^{-1}(c)$$

with the property that the projection of $\mathscr{C}_n$ onto D_n, that is, the set

$$\left\{\mathbf{x}; \quad \mathbf{x} \in \{0,1\}^{|D_n|}, \quad \exists \mathbf{y} \in \mathscr{C}_n \qquad y_i = x_i, \quad \forall i \in D_n\right\},$$

is the whole $\{0,1\}^{|D_n|}$; here the function $h^{-1} : [0,1] \to [0, \frac{1}{2}]$ is the inverse of the binary entropy function (which exists when the function h is restricted to the left half of the unit interval). Our result will immediately follow from the claim that

$$1 - h^{-1}(c) \geqslant c. \tag{2.1}$$

In order to verify this claim suppose to the contrary that

$$1 - h^{-1}(c) < c. \tag{2.2}$$

Now let $m(\mathbf{z})$ denote, for every $\mathbf{z} \in \{0,1\}^{|\overline{D_n}|}$, the number of those elements of $\mathscr{C}_n$ whose projection onto its coordinates in $\overline{D_n}$ equals $\mathbf{z}$. Obviously,

$$|\mathscr{C}_n| = \sum_{\mathbf{z}\in\{0,1\}^{|\overline{D_n}|}} m(\mathbf{z}) \leqslant \exp_2\left(n - |D_n|\right) \max_{\mathbf{z}\in\{0,1\}^{|\overline{D_n}|}} m(\mathbf{z}).$$

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

This together with our hypothesis (2.2) implies, in particular, that, for all sufficiently large n,

$$\max_{\mathbf{z}\in\{0,1\}^{|\overline{D_n}|}} m(\mathbf{z}) > 4.$$

Let n_0 be the threshold above which the last inequality holds, and consider a fixed $\mathbf{z}$ with $m(\mathbf{z}) > 4$. Further, let $\mathscr{E}'$ be the set of those sequences in $\mathscr{C}_n$ whose projection onto $\overline{D_n}$ is this $\mathbf{z}$. Since $\mathscr{E}'$ has at least 5 elements, for any fixed coordinate $j \in D_n$ there must be 3 of them, $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$, say, which are all equal in their jth coordinate. Fix such a coordinate j, and let $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$ be the corresponding strings. Finally, consider the binary sequence $\mathbf{a} = a_{i_1}a_{i_2}\ldots a_{i_{|D_n|}}$ defined by setting, for every $i \in D_n$ the coordinate a_i equal to the only value from $\{0,1\}$ which occurs twice or never among the ith coordinates of $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$. This guarantees the existence of a sequence $\mathbf{v} \in \mathscr{C}_n$, different from $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$ (since its jth coordinate differs from that of $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$), whose projection onto D_n is precisely this $\mathbf{a}$. But then the 4 sequences of the 'string quartet' $\mathbf{v}$, $\mathbf{w}$, $\mathbf{x}$ and $\mathbf{y}$ do not satisfy the criterion that both 0 and 1 occur twice in some of their coordinates, since in every coordinate i belonging to D_n the corresponding value a_i appears among the four ith coordinates an odd number of times, while in all of the remaining coordinates the common value of the three original sequences appears at least 3 times among the 4 sequences. This contradiction shows the impossibility of (2.2) and thus we must have (2.1). An easy calculation shows (2.1) to be equivalent to the relation

$$c \leqslant h(1-c).$$

The largest value for which this holds satisfies the equation

$$c = h(1-c),$$

whose solution gives our upper bound. □

3. Proof of Theorem 1.2

Once again, the proof of the lower bound is simple, and proceeds as follows. Let F be a random family of N binary strings of length n, where each member of F is chosen, independently, by letting each of its coordinates, randomly and independently, be 1 with probability p and 0 with probability $1-p$, where p will be chosen later. The expected number of fourtuples in F for which there is no coordinate in which exactly one of them has a 1 is $\binom{N}{4}(1-4p(1-p)^3)^n$. The expected number of pairs of identical strings in F is $\binom{N}{2}(p^2+(1-p)^2)^n$. Therefore, there is a choice of N strings in which the number of forbidden fourtuples plus the number of forbidden pairs is at most $\binom{N}{4}(1-4p(1-p)^3)^n + \binom{N}{2}(p^2+(1-p)^2)^n$. By deleting an arbitrarily chosen string from each such forbidden configuration we conclude that, for every p between 0 and 1 and for every N,

$$M(n,C) \geqslant N - \binom{N}{4}(1-4p(1-p)^3)^n - \binom{N}{2}(p^2+(1-p)^2)^n.$$

Choosing $p = 1/4$ and $N = \lfloor (64/37)^{n/3} \rfloor$, it follows that $M(n,C) \geqslant c(64/37)^{n/3}$ for some absolute positive constant c, supplying the desired lower bound.

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

We next give a detailed proof of the upper bound. Suppose first that the set $\mathscr{C}'_n \subseteq \{0,1\}^n$ achieves the maximum in the definition of $M(n, C)$, and let us write

$$d = \limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}'_n|.$$

Clearly, then there exists, for every n, a set $\mathscr{C}_n \subseteq \mathscr{C}'_n$ such that all the strings in $\mathscr{C}_n$ have the same number, say $np(n)$, coordinates equal to 1, while we continue to have

$$d = \limsup_{n\to\infty} \frac{1}{n} \log |\mathscr{C}_n|. \tag{3.1}$$

Without restricting generality we can suppose the existence of the limit

$$\lim_{n\to\infty} p(n) = q. \tag{3.2}$$

(In order to verify the existence of $\mathscr{C}_n$ as above, notice that $\mathscr{C}'_n$ can be partitioned into at most $n+1$ subsets with the property that, within each of them, the number of the 1 s in every string is the same. Choosing $\mathscr{C}_n$ as one of these classes with maximum cardinality, the statement follows.)

Let us fix n and $\mathscr{C}_n$ as above and write $p = p(n)$. Further, let us denote by $p_i = p_i(n)$ the fraction of those elements of $\mathscr{C}_n$ whose ith coordinate equals 1. Clearly, we must have

$$\frac{1}{n} \sum_{i=1}^{n} p_i = p. \tag{3.3}$$

Next we claim that, for every ϵ and n sufficiently large,

$$\frac{1}{n} \log \binom{|\mathscr{C}_n|}{2} \leqslant h(2p(1-p)) + \epsilon. \tag{3.4}$$

In order to prove this inequality, let us first define, for every $i \in [n]$, the function $f_i : \binom{\mathscr{C}_n}{2} \to \{0,1\}$ by setting $f_i(A) = 1$ if and only if the two binary strings forming A differ in their ith coordinates and set $f_i(A) = 0$ otherwise.

We claim that, if $A \neq B$ for some $A \in \binom{|\mathscr{C}_n|}{2}$, $B \in \binom{|\mathscr{C}_n|}{2}$, then there exists an $i \in [n]$ for which

$$f_i(A) \neq f_i(B). \tag{3.5}$$

This is perfectly clear if $A \cap B \neq \emptyset$, and is a consequence of our hypothesis on $\mathscr{C}_n$ otherwise. As a matter of fact, if A and B are disjoint, then their union defines a 'string quartet' at least one of whose coordinates, say, the ith, is in C. But then, necessarily, no matter how A and B divide the quartet into two couples, we have (3.5).

Let X^n be a random variable uniformly distributed over the unordered pairs of distinct elements of $\mathscr{C}_n$. We define, for every $i \in [n]$, the random variable Z_i by setting $Z_i = f_i(X^n)$. We have

$$H(X^n) = \log \binom{|\mathscr{C}_n|}{2}$$

and

$$H(Z_i) = h\left(2p_i(1-p_i)\frac{|\mathscr{C}_n|}{|\mathscr{C}_n|-1}\right) < h(2p_i(1-p_i)) + \epsilon.$$

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

Since by its definition $Z^n = Z_1 \dots Z_n$ is a function of X^n , we have $H(Z^n) \leqslant H(X^n)$. On the other hand (3.5) implies that the function defining Z^n from X^n is injective, whence also Z^n determines X^n and thus

$$H(X^n) = H(Z^n).$$

Comparing this with the above and applying the well-known subadditivity of entropy (see [2]), we obtain

$$\log \binom{|\mathscr{C}_n|}{2} = H(X^n) \leqslant \sum_{i=1}^{n} H(Z_i) = \sum_{i=1}^{n} \big[h(2p_i(1-p_i)) + \epsilon\big] \leqslant n\big[h(2p(1-p)) + \epsilon\big], \quad (3.6)$$

where the rightmost inequality follows by the easily checkable cap-convexity of the function $h(2x(1-x))$ in x, in virtue of (3.3); this establishes (3.4).

We shall use (3.4) in combination with another inequality we now intend to prove. In fact, we claim that

$$\frac{1}{n} \log |\mathscr{C}_n| \lesssim \frac{1}{2-q} \sum_{i=1}^{n} (1-p_i^2) h\left(\frac{2p_i}{1+p_i}\right). \quad (3.7)$$

This is the core of our proof and it will take some time before we complete its verification. We will use the concept of graph entropy introduced in [8] (see also [9]). Let the graph $G = G_n$ have vertex set

$$V(G) = \binom{\mathscr{C}_n}{2},$$

and let the vertices $A \in \binom{\mathscr{C}_n}{2}$, $B \in \binom{\mathscr{C}_n}{2}$ be adjacent in G if there is a coordinate $i \in [n]$ for which exactly one of the (3 or 4) sequences in $A \cup B$ is equal to 1. Further, let P be the uniform probability distribution on $V(G)$.

We will need appropriate lower and upper bounds on $H(G, P)$, the entropy of the graph G with respect to the distribution P. We recall that graph entropy is formally defined as

$$H(G, P) = \min_{X \in Y \in S(G),\, P_X = P} I(X \wedge Y),$$

where $S(G)$ denotes the family of the stable sets of vertices in G. A subset of the vertex set is called stable if it does not contain any edge. We recall that the mutual information $I(X \wedge Y)$ of the random variables X and Y equals $H(X) + H(Y) - H(X, Y)$, where, for instance, $H(X, Y)$ is the entropy of the random variable (X, Y). It is immediate from this definition that, if K is a complete graph and P an arbitrary distribution on its vertex set, then $H(K, P) = H(P)$, and thus graph entropy enriches our possibilities for obtaining entropy-based bounds.

First let us prove that, for any $\epsilon > 0$ and sufficiently large n, the graph G just defined on $\mathscr{C}_n$ satisfies

$$\log \alpha(G) \leqslant q \log |\mathscr{C}_n| + n\epsilon \quad (3.8)$$

where, following tradition, we denote by $\alpha(G)$ the cardinality of the largest stable set in $V(G)$. Now let $\mathscr{S} \subseteq V(G)$ be a stable set of G of maximum cardinality. Actually, since, by our hypothesis, if $A \in \binom{\mathscr{C}_n}{2}$, $B \in \binom{\mathscr{C}_n}{2}$, $|A \cup B| = 4$, then there exists an $i \in [n]$ such that

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

exactly one of the four coordinates of these 4 sequences is equal to 1, such disjoint sets A and B cannot both belong to $\mathscr{S}$. In other words, we see that the vertices of any stable set, and thus of $\mathscr{S}$, form an intersecting family in $\binom{\mathscr{C}_n}{2}$. Since we want to prove (3.8), we can suppose $|\mathscr{S}| > 3$. But then there exists an $\mathbf{x} \in \mathscr{C}_n$ such that $A \in \mathscr{S}$ implies $\mathbf{x} \in A$. Consider

$$\hat{\mathscr{S}} = \{\mathbf{y}; \quad \{\mathbf{x}, \mathbf{y}\} \in \mathscr{S}\}$$

and

$$D = \{i; x_i = 1\}.$$

Clearly, if $\mathbf{y} \neq \mathbf{z}$ for $\mathbf{y} \in \hat{\mathscr{S}}$, $\mathbf{z} \in \hat{\mathscr{S}}$, then these two sequences cannot differ in any coordinate $i \in \overline{D}$. This implies that, given any string quartet $\mathbf{y}, \mathbf{y}', \mathbf{y}'', \mathbf{y}'''$ with all its strings from $\hat{\mathscr{S}}$, there cannot be any coordinate $i \in \overline{D}$ in which these 4 strings have just one 1. But then these 4 strings that are all from $\mathscr{C}_n$ must have such a coordinate $j \in D$. Let $\mathscr{C}_n^* \subseteq \{0,1\}^{|D|}$ denote the set of projections of the strings in $\hat{\mathscr{S}}$ onto their coordinates in D. We have, for every $\epsilon > 0$ and sufficiently large n, by the definition (3.1) of d that

$$|\mathscr{S}| - 1 = |\hat{\mathscr{S}}| = |\mathscr{C}_n^*| \leqslant \exp\left[|D|\left(d + \frac{\epsilon}{4}\right)\right] \leqslant \exp\left[nq\left(d + \frac{\epsilon}{4}\right)\right] \leqslant \exp(n\epsilon) \cdot |\mathscr{C}_n|^q. \quad (3.9)$$

The last inequality yields (3.8), which in turn implies

$$H(G,P) \geqslant H(P) - \log \alpha(G) = \log \left|\binom{\mathscr{C}_n}{2}\right| - \log \alpha(G) \geqslant (2-q)\log|\mathscr{C}_n| - n\epsilon, \quad (3.10)$$

where the first inequality, an easy consequence of the definition of graph entropy, is explained more in detail in [9, p. 568].

In order to complete the verification of (3.7) we turn to the proof of our upper bound on $H(G,P)$. Given any $i \in [n]$, let G_i be the graph having the same vertex set as G and an edge set $E(G_i) \subseteq E(G)$ defined by making the vertices $A \in \binom{\mathscr{C}_n}{2}$, $B \in \binom{\mathscr{C}_n}{2}$ adjacent in G_i if exactly one of the (3 or 4) sequences in $A \cup B$ has its ith coordinate equal to 1. Since every couple of sets $\{A, B\}$ must satisfy this for at least one $i \in [n]$, we immediately see that

$$G \subseteq \bigcup_{i=1}^{n} G_i$$

(where for two graphs F and G on the same vertex set V $F \cup G$ denotes the graph on V with edge set $E(F) \cup E(G)$). It follows from the sub-additivity of graph entropy (see [9, Corollary 1, p. 562]) that

$$H(G,P) \leqslant \sum_{i=1}^{n} H(G_i, P). \quad (3.11)$$

Next we want to check that for every $i \in [n]$ we have

$$H(G_i, P) \leqslant (1 - p_i^2) h\left(\frac{2p_i}{1+p_i}\right). \quad (3.12)$$

To see this, introduce the graph F with vertex set $V(F) = \{0,1,2\}$ and the single edge $\{0,2\}$. Observe that the function $g_i : \binom{\mathscr{C}_n}{2} \to \{0,1,2\}$, defined by setting $g_i(\{\mathbf{x},\mathbf{y}\}) = 0$ if $x_i = y_i = 0$, $g_i(\{\mathbf{x},\mathbf{y}\}) = 1$ if $x_i = y_i = 1$ and $g_i(\{\mathbf{x},\mathbf{y}\}) = 2$ otherwise, acts on the vertices of

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

G_i in an edge-preserving manner. Next consider the probability distribution P_i on $\{0,1,2\}$ defined by

$$P_i(t) = P(g_i^{-1}(t)) = \sum_{A \in \binom{\mathscr{C}_n}{2}, g_i(A)=t} P(A).$$

Clearly, in the limit of n going to infinity, when the effect of not allowing repetitions can be neglected, we can suppose that $P_i(0) = (1-p_i)^2$, $P_i(1) = p_i^2$ and $P_i(2) = 2p_i(1-p_i)$. Thus, as an easy consequence of the above definition of graph entropy, one sees that

$$H(G_i, P) = H(F, P_i) = (1-p_i^2)h\left(\frac{2p_i}{1+p_i}\right).$$

Now, from (3.11) and (3.12) we get

$$H(G,P) \leqslant \sum_{i=1}^{n} (1-p_i^2)h\left(\frac{2p_i}{1+p_i}\right) \tag{3.13}$$

right away. Let the function $l : [0,1] \to [0,1]$ be the upper convex envelope of the function $k(t) = (1-t^2)h(\frac{2t}{1+t})$, $t \in [0,1]$. With this notation, (3.13) and (3.3) imply $H(G,P) \leqslant nl(p)$, whence, in virtue of the definition (3.2) of q, for every $\epsilon > 0$ and sufficiently large n one has

$$H(G,P) \leqslant n(l(q)+\epsilon). \tag{3.14}$$

Comparing this and (3.10) we obtain, for every $\epsilon > 0$ and sufficiently large n,

$$\frac{1}{n}\log|\mathscr{C}_n| \leqslant \frac{l(q)+\epsilon}{2-q} + \epsilon. \tag{3.15}$$

This inequality together with (3.4) gives that, for every ϵ and sufficiently large n,

$$\frac{1}{n}\log|\mathscr{C}_n| \leqslant \max_{q\in[0,1]} \min\left\{\frac{h(2q(1-q))}{2}, \quad \frac{l(q)+\epsilon}{2-q}\right\} + \epsilon,$$

whereby the definition of d yields

$$d \leqslant \max_{q\in[0,1]} \min\left\{\frac{h(2q(1-q))}{2}, \quad \frac{l(q)}{2-q}\right\}. \tag{3.16}$$

An easy calculation shows that the right-hand side of this is strictly less than $1/2$ as claimed; in fact it suffices to verify that if $q \neq 1/2$ then $\frac{h(2q(1-q))}{2} < 1/2$, while if $q = 1/2$ we have $\frac{l(q)}{2-q} < 1/2$. □

4. Remarks

The determination of the values of $M(n,A)$ and $M(n,C)$, at least in an asymptotic sense, seems to be the natural analogue for fourtuples of perhaps one of the most difficult problems in extremal set theory, namely that of deciding whether the maximum number $F(n,3)$ of binary sequences of length n without 3 of them forming a strong Δ-system satisfies $\limsup_{n\to\infty} \frac{1}{n}\log F(n,3) < 1$. A recent survey article of Kostochka on Δ-systems [12] shows how far we still are from answering this question. It is therefore significant that for 'string quartets' much more could be said on analogous questions. One of the

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

reasons for this is that by considering quartets as pairs of pairs we can introduce more structure into our analysis.

Combining our methods here with an additional (simple) argument we can obtain results concerning some more general analogues of the Δ-system question. For $k \geqslant m \geqslant 0$, and n satisfying $2^{n-1} > k$, let $f(n,k,m)$ denote the maximum possible number of binary strings of length n such that for any k of them there is a coordinate in which exactly m of them have a 1. In this notation, Theorem 1.1 supplies bounds for $M(n,A) = f(n,4,2)$, whereas Theorem 1.2 provides bounds for $f(n,4,1)$. By exchanging the roles of 0 and 1 it follows that $f(n,k,m) = f(n,k,k-m)$, and trivially $f(n,1,0) = f(n,1,1) = 2^n - 1$, $f(n,2,1) = 2^n$, and, for every $k \geqslant 2$, $f(n,k,0) = f(n,k,k) = 2^{n-1}$. The Δ-system question is that of determining if $f(n,3,1) = f(n,3,2)$ is at most $(2-\epsilon)^n$ for some fixed positive ϵ. Our methods here suffice to prove the following.

Theorem 4.1. *There exists a positive constant δ such that, for every fixed $k \geqslant 4$ and for every m satisfying $k > m > 0$, $f(n,k,m) \leqslant (2-\delta)^n$.*

Proof. If m is neither $k-1$ nor 1, we can apply the method in the proof of Theorem 1.1 to show that in this case

$$\limsup_{n\to\infty} \frac{1}{n} \log f(n,k,m) \leqslant 0.773\ldots. \tag{4.1}$$

Indeed, given a family C of strings of length n, whose cardinality is at least 2^{cn} with $c + h^{-1}(c) > 1$ and n sufficiently large, we split it into two disjoint sub-families C_1 and C_2 of nearly equal cardinalities, and apply Sauer's lemma to one of them, say C_1, to get a large set D_n of coordinates on which C_1 has all possible projections. Then we apply the pigeonhole principle to the projections of the strings in C_2 on the rest of the coordinates, to get $k-1$ members $\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}$ of C_2 having identical projections on the coordinates not in D_n. Now we can add to these $k-1$ strings a string $\mathbf{x}_k$ from C_1 whose projection on D_n is chosen to ensure that, for each coordinate in D_n, the number of 1 s the strings $\mathbf{x}_1, \ldots, \mathbf{x}_k$ have differs from m. The number of 1 s these strings have in each coordinate not in D_n is clearly either 0 or 1 or $k-1$ or k, which are all different from m. This proves (4.1).

It remains to consider the case $m = 1$ (which is equivalent to the case $m = k-1$, as $f(n,k,1) = f(n,k,k-1)$). We show that in this case

$$\limsup_{n\to\infty} \frac{1}{n} \log f(n,k,m) \leqslant \frac{1}{2} \log 3.$$

Indeed, recall that $k \geqslant 4$, put $t = \lceil k/2 \rceil$, let C be a family of binary strings of length n, and suppose that $\binom{|C|}{2} > (t-1)3^n$. Define $S = \{\mathbf{x} + \mathbf{x}' : \mathbf{x}, \mathbf{x}' \in C, \mathbf{x} \neq \mathbf{x}'\}$, where the sum is the usual sum of vectors (over the integers). By the pigeonhole principle there are t equal members of S , corresponding to t pairs of strings in C, with equal sums. Trivially these pairs are disjoint. If k is even, all the strings in these pairs are k strings in which the number of 1 s in each coordinate is either 0 or $t = k/2$ or k, which are all different from $m = 1$, as needed. If k is odd, then $t \geqslant 3$. In this case simply omit an arbitrarily chosen string from the $2t$ strings in these pairs, obtaining a set of k strings in which the

https://doi.org/10.1017/S0963548300004375 Published online by Cambridge University Press

number of 1 s in each coordinate lies in the set $\{0, t, t-1, k\}$, which does not contain 1. This completes the proof. □

By being more careful we can improve the above general upper bound for $f(n,k,m)$ in various cases. In particular, we can show that $f(n,2t,t) \leqslant 2^{\epsilon(t)n}$, where $\epsilon(t)$ tends to 0 as t tends to infinity. We omit the details, and intend to return to these questions in a subsequent paper.

Acknowledgement

Part of this work was done during a visit by the first author to the University of Rome, 'La Sapienza', and he would like to thank his hosts for their hospitality.

References

[1] Alon, N. and Spencer, J. (1992) *The Probabilistic Method*, Wiley, New York.
[2] Csiszár, I. and Körner, J. (1981) *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York (1982) and Akadémiai Kiadó, Budapest.
[3] Dyachkov, A. G. and Rykov, V. V. (1982) Bounds on the length of disjunctive codes. *Probl. Per. Inf.* **18** 7–13. (In Russian.)
[4] Erdős, P. and Szemerédi, E. (1978) Combinatorial properties of systems of sets. *J. Combin. Theory Ser. A* **24** 308–313.
[5] Erdős, P., Frankl, P. and Füredi, Z. (1982) Families of finite sets in which no set is covered by the union of two others. *J. Combin. Theory Ser. A* **33** 158–166.
[6] Frankl, P. and Füredi, Z. (1984) Union-free hypergraphs and probability theory. *European J. Combin.* **5** 127–131.
[7] Kleitman, D. and Spencer, J. (1973) Families of k-independent sets. *Discrete Math.* **6** 255–262.
[8] Körner, J. (1973) Coding of an information source having ambiguous alphabet and the entropy of graphs. *Trans. 6th Prague Conference on Inform. Theory, etc., 1971*, Academia, Prague, pp. 411–425.
[9] Körner, J. (1986) Fredman–Komlós bounds and information theory. *SIAM J. Alg. Discrete Meth.* **7** 560–570.
[10] Körner, J. and Orlitsky, A. (1998) Zero-error information theory. *IEEE Trans. Inform. Theory* **44** 2207–2229.
[11] Körner, J. and Simonyi, G. (1988) Separating partition systems and locally different sequences. *SIAM J. Discrete Math.* **1** 355–359.
[12] Kostochka, A. V. (1998) Extremal problems on Δ-systems. Manuscript.
[13] Lindström, B. (1969) Determination of two vectors from the sum. *J. Combin. Theory Ser. A* **6** 402–407.
[14] Sauer, N., (1972) On the density of families of sets. *J. Combin. Theory Ser. A* **13** 145–147.
[15] Shearer, J. B. (1996) On cancellative families of sets. *Electronic J. Combinatorics* **1**.
[16] Shelah, S. (1972) A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific J. Math.* **41** 247–261.
[17] Sperner, E. (1928) Ein Satz über Untermengen einer endlichen Menge. *Math. Zeitschr.* **27** 544–548.
[18] Tolhuizen, L. (1999) New rate pairs in the zero-error capacity region of the binary multiplying channel without feedback. Submitted to *IEEE Trans. Inform. Theory*.
[19] Vapnik, V. N. and Chervonenkis, A. Y. (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.