Erasure list-decodable codes and Turán hypercube problems

Noga Alon [*]

Abstract

We observe that several vertex Turán type problems for the hypercube that received a considerable amount of attention in the combinatorial community are equivalent to questions about erasure list-decodable codes. Analyzing a recent construction of Ellis, Ivan and Leader, and determining the Turán density of certain hypergraph augmentations we obtain improved bounds for some of these problems.

AMS classification: 05C35, 94B65
Keywords: Erasure codes, list decodable codes, Turán problems, hypercube

1 Introduction and results

1.1 Erasure codes and Turán hypercube problems

A set $\mathcal{C}$ of binary vectors of length n is a (d, L)-list decodable erasure code of length n (a (d, L, n)-code, for short) if for every codeword w, after erasing any d-bits of w, the remaining part of the vector has at most L possible completions into codewords of $\mathcal{C}$. Erasure list-decodable codes are considered in [6], see also [2] and the references therein. These papers deal with codes of rate smaller than 1, that is, the cardinality of $\mathcal{C}$ is exponentially smaller than 2^n.

Here we consider much denser codes, where the cardinality of $\mathcal{C}$ is a constant fraction of all 2^n vectors. This range of the parameters is not very natural from the information theoretic point of view, but it is equivalent to a problem that received a considerable amount of attention in the combinatorial community, see [11], [8], [1], [7], [3], [10], [5], [4]. Indeed, $\mathcal{C}$ is a (d, L, n)-code if and only if it is a subset of vertices of the discrete n-cube Q_n that contains at most L vertices of any d-dimensional subcube of Q_n. In this language, for example, the result of [11], proved independently in [8], is that the maximum possible cardinality of a $(2, 3, n)$-code is $\lceil 2^{n+1}/3 \rceil$.

[*] Princeton University, Princeton, NJ, USA and Tel Aviv University, Tel Aviv, Israel. Email: nalon@math.princeton.edu. Research supported in part by NSF grant DMS-2154082.

1.2 $(d, 2^d-1, n)$-codes

An intriguing special case of the general problem of determining or estimating the maximum possible cardinality of (d, L, n)-codes is the cases $L = 2^d - 1$ corresponding to codes $\mathcal{C}$ that contain no full copy of a d-subcube. Here it is more natural to consider the complement and denote by $g(n, d)$ the smallest cardinality of a subset of the vertices intersecting every d-subcube. Let γ_d denote the limit $\lim_{n \mapsto \infty} g(n, d)/2^n$ (it is easy to see that the limit exists as for any fixed d, $g(n, d)/2^n$ is a monotone increasing function of n). Trivially, $\gamma_1 = 1/2$, and the result of [11] and [8] mentioned above is that $\gamma_2 = 1/3$. In [1] it is shown that $\gamma_d \geq \log_2(d+2)/2^{d+2}$. It has been a folklore conjecture (see [3]) that $\gamma_d = 1/(d+1)$ but this is refuted in a very strong sense in a recent paper of Ellis, Ivan and Leader [4], where it is shown that $(\gamma_d)^{1/d} \leq 2^{-1/8+o(1)}$. As mentioned in [4] we observed that their argument can be improved to show that $(\gamma_d)^{1/d} \leq 2^{-1/2+o(1)}$. This is stated in the following proposition.

Proposition 1.1. *For every large k and every n there is a subset of less than a fraction of 2^{-k} of the vertices of the n-cube that intersects the set of vertices of any cube of dimension $d = 2k + 3\log_2 k$.*

1.3 Codes of positive density

Another range of the parameters of (d, L, n)-codes that has been studied in quite a few combinatorial papers deals with the minimum possible $L = L(d)$ so that there exists infinitely many (d, L, n)-codes of positive density. More precisely, let $L(d)$ denote the smallest possible L so that there exists an $\varepsilon = \varepsilon(d) > 0$ such that for every n there is a (d, L, n)-code of cardinality at least $\varepsilon 2^n$. The problem of determining or estimating $d(L)$ is considered in [10] (where it is denoted by $\mu(d)$.) A conjecture suggested in [3] asserts that $L(d) = \binom{d}{\lfloor d/2 \rfloor}$. It is easy to see that this is always an upper bound for $L(d)$, and that it holds as equality for $d \leq 3$. However, somewhat surprisingly this conjecture too is refuted by the recent construction of [4] which shows that $L(d)$ is at most $(5/6)\binom{d}{\lfloor d/2 \rfloor}$ for every $d \geq 4$. The authors of [10] proved a lower bound for $L(d)$, showing that it is at least $t_2(d) + t_3(d)$, where $t_2(d)$ is 0 if $\lceil d/3 \rceil$ is odd and 1 otherwise, and $t_3(d)$ is $3^{d/3}$ for $d \equiv 0 \bmod 3$, is $4 \cdot 3^{(d-4)/3}$ for $d \equiv 1 \bmod 3$ and is $2 \cdot 3^{(d-2)/3}$ for $3 \equiv 2 \bmod 3$. In particular, this shows that

$$L(5) \geq 7, L(6) \geq 10, L(7) \geq 12, L(8) \geq 18, L(9) \geq 27, L(10) \geq 37.$$

Here we improve the lower bounds for all $d \geq 5$, proving, in particular, the following

Proposition 1.2.

$$L(5) \geq 8, L(6) \geq 12, L(7) \geq 20, L(8) \geq 32, L(9) \geq 48, L(10) \geq 80.$$

For large d we prove that $L(d) \geq d \cdot 3^{(d-6)/3}$ for d divisible by 3 and obtain a similar bound for d that is not divisible by 3. The improved lower bounds are obtained by applying the simple result about graph and hypergraph augmentations described in the following subsection.

We also improve the upper bounds as follows:

Theorem 1.3.

$$L(5) = 8, L(6) \leq 16, L(7) \leq 28$$

and for large d,

$$L(d) \leq (c + o(1)) \binom{d}{\lfloor d/2 \rfloor}$$

where

$$c = \lim_{t \mapsto \infty} \frac{(2^t - 2)(2^t - 4) \ldots (2^t - 2^{t-1})}{(2^t - 1)^{t-1}}$$

is roughly 0.29.

Note that by the results above the exact values of $L(d)$ for $1 \leq d \leq 5$ are given by the sequence $1, 2, 3, 5, 8$.

1.4 Graph and hypergraph augmentations

For a graph $G = (V, E)$ and an integer $r \geq 2$, let the r-augmentation of G, denoted by $G(r)$, be the r-uniform hypergraph $(V \cup S, \{e \cup S : e \in E\})$, where $S \cap V = \emptyset$ and $|S| = r - 2$. Thus $G(r)$ is obtained from G by adding the same set of $r - 2$ vertices to each edge of G. This set is called the stem of $G(r)$. More generally, for a k-uniform hypergraph $H = (V, E)$ and an integer $r \geq k$, let the r-augmentation of H, denoted $H(r)$, be the r-uniform hypergraph $(V \cup S, \{e \cup S : e \in E\})$, where $S \cap V = \emptyset$ and $|S| = r - k$.

For a fixed r-uniform hypergraph F and for an integer n let $ex(n, F)$ denote the maximum possible number of edges in an r-uniform hypergraph on n vertices that contains no copy of F. The Turán density $\pi(F)$ of F is the limit, as n tends to infinity, of the ratio $ex(n, F)/\binom{n}{r}$. It is easy and well known that this limit always exists, and lies in $[0, 1]$. Indeed the expected number of edges on a random set of $n - 1$ vertices among the n vertices of an F-free r-uniform hypergraph with $ex(n, F)$ edges is

$$\frac{ex(n, F)\binom{n-1}{r}}{\binom{n}{r}}.$$

Therefore this quantity is a lower bound for $ex(n - 1, F)$ implying that the ratio

$$\frac{ex(n, F)}{\binom{n}{r}}$$

is a monotone decreasing function of n. Therefore the limit exists (and is obviously in $[0,1]$.)

The recent construction of Ellis, Ivan and Leader in [4] implies that if the chromatic number of a graph G satisfies $\chi(G) \geq 4$, then the Turán density of $G(r)$ is at least 0.29 for every r. (The construction in [4] is described for $G = K_4$, but it is not difficult to check that it works for every graph G of chromatic number at least 4).

Here we observe that if $\chi(G) \leq 3$ then the Turán density of $G(r)$ tends to zero as r tends to infinity. This gives a full characterization of all the fixed graphs G that must appear as links in any r-uniform hypergraph of positive density (with at least $2r+1$ vertices, say), provided r is sufficiently large. This result has been proved independently by Robert Johnson [9]

Proposition 1.4. *For every fixed graph G with chromatic number at most 3, the limit of the Turán density of $G(r)$ as r tends to infinity is 0.*

The argument easily extends to augmentations of hypergraphs, giving the following

Proposition 1.5. *For any fixed k-uniform hypergraph H in which the set of vertices is the disjoint union of $k+1$ subsets, so that every edge contains at most one vertex in each subset, the Turán density of $H(r)$ tends to 0 as r tends to infinity*

Remark: By averaging over r, Proposition 1.4 implies that for every fixed $\varepsilon > 0$ and every fixed graph G of chromatic number at most 3, if $n > n_0(G, \varepsilon)$ then any family of at least $\varepsilon 2^n$ subsets of $[n] = \{1, 2, \ldots, n\}$ contains a copy of $G(r)$ for some r. Here, too, the construction in [4] implies that this is false for graphs G of chromatic number at least 4. Similarly, Proposition 1.5 implies the corresponding result for the hypercube.

2 Proofs

2.1 Augmentations

In this subsection we describe the short proof of Proposition 1.4. The proof of Proposition 1.5 is essentially identical.

Fix an $\varepsilon > 0$, suppose $n \geq 2r+1$ and let H be an r-uniform hypergraph on n vertices with at least $\varepsilon\binom{n}{r}$ edges. By averaging there is a subset U of $2r+1$ vertices so that H contains at least $\varepsilon\binom{2r+1}{r}$ edges in U. Let W be a random subset of size $r+1$ of U. The expected number of edges contained in W is at least $\varepsilon(r+1)$. If W contains $k \geq 3$ edges, then any collection of 3 of them gives a copy of $K_3(r)$. Thus we get $\binom{k}{3}$ such copies on the set of vertices W. By convexity (assuming, say, $\varepsilon(r+1) > 10$) this implies that the total number of copies of $K_3(r)$ that are contained in U is at least

$$\binom{2r+1}{r+1} \cdot \binom{\varepsilon(r+1)}{3}.$$

By averaging over the $\binom{2r+1}{r-2}$ possible stems we get that there is one common stem for at least

$$\frac{\binom{2r+1}{r+1}}{\binom{2r+1}{r-2}} \cdot \binom{\varepsilon(r+1)}{3} > \frac{\varepsilon^3}{2}\binom{r}{3}$$

copies of $K_3(r)$, where here we assumed, say, $r > 10/\varepsilon$. This gives the existence of a graph F on a subset of $r+3$ vertices of U so that F contains more than $\frac{\varepsilon^3}{2}\binom{r}{3}$ triangles and our hypergraph contains a copy of $F(r)$. By the known results about the Turán density of 3-uniform, 3-partite hypergraphs first proved in [12], for every s and every sufficiently large $r > r_0(\varepsilon, s)$, F contains a complete 3-partite graph T with s vertices in each vertex class. Since $F(r)$ contains $T(r)$ this completes the proof of the proposition. □

2.2 Hitting subcubes

In this subsection we describe the proof of Proposition 1.1. The proof is identical to the one in [4] with one modification, replacing a naive estimate for the maximum possible number of k-wise independent vectors in F_2^s by the Plotkin bound [15], which is a classical result in the theory of Error Correcting Codes.

For simplicity we omit all floor and ceiling signs whenever these are not crucial. All logarithms are in base 2 unless otherwise specified.

Following the notation in [4], for integers $t > s$ and $r > s$, let $D_r(s,t)$ denote the r-uniform hypergraph obtained by adding a stem of size $r-s$ to every edge of the complete s-uniform hypergraph T on t vertices. In the notation of the previous subsection $D_r(s,t)$ is $T(r)$. In [4], Theorem 6, it is proved that for every fixed k and every (large) r, the Turán density of $D_r(k, 8k+1)$ is at least $1 - O(2^{-k})$. The following lemma provides a quantitative improvement.

Lemma 2.1. *For every fixed (large) integer k and every (large) r, the Turán density of $D_r(k+2\log k, 2k+3\log k)$ is larger than $1-2^{-k}$.*

Proof. Suppose r is large and consider the hypergraph H on a set of $2^{r+k}-1$ vertices indexed by the nonzero vectors in F_2^{r+k}, where an r-set forms an edge iff it is linearly independent. It is easy to see that the density of this hypergraph is larger than $1-2^{-k}$. We claim that it contains no copy of $D_r(s,t)$ where $s = k+2\log k, t = 2k+3\log k$. Indeed, as in the proof in [4], the existence of such a set would give a collection of t binary vectors in F_2^{k+s} so that every subset of s of them is linearly independent. For completeness we describe the argument. Let $v_1, v_2, \ldots, v_{r-s} \in F_2^{r+k}$ be the vectors corresponding to the $(r-s)$ vertices of the stem of a copy of $D = D_r(s,t)$ and let $u_1, u_2, \ldots, u_t \in F_2^{r+k}$ be the vectors corresponding to the other t vertices of D. Then the vectors v_i are linearly independent and span a subspace V of dimension $r-s$. The space F_2^{r+k}/V is isomorphic to F_2^{k+s} and for each subset of s of the vectors u_i, the cosets $u_i + V$ are linearly independent in this space, providing the desired collection of t vectors.

Let A be the $k+s$ by t matrix whose columns are these t vectors and consider the linear code whose parity check matrix is A. This is the code consisting of all binary vectors of length t that are orthogonal to every row of A. The dimension of this code is at least $t-(k+s) = \log k$ and hence the number of vectors in it is at least k. However, the minimum distance of this code is at least $s+1$, since every set of s columns of A is linearly independent. By the Plotkin bound it follows that the number of vectors in the code cannot exceed $2\frac{(s+2)}{2s+2-t} < k$, contradiction. Therefore this hypergraph contains no copy of $D_r(s,t)$. The assertion of the lemma follows by considering blow ups of this hypergraph, which for large r hardly change the density. Here is an explicit description of this blow up. For $n \geq 2^{r+k}-1$ take, for each vertex w of H, a set of vertices X_w of either $\lfloor n/(2^{r+k}-1)\rfloor$ or $\lceil n/(2^{r+k}-1)\rceil$ copies of w, so that the total number of vertices is n. A set of r vertices forms an edge iff they are copies of r distinct vertices of H that form an edge of H. Since no edge contains more than a single vertex from each fixed set X_w, any subgraph of this new hypergraph is also a subgraph of H, and thus the new hypergraph contains no copy of $D_r(s,t)$. Since $2^{r+k}-1$ is much bigger than r^2, almost every set of r vertices of the new hypergraph contains at most one vertex from each set X_w, implying that the density of the new hypergraph is at least that of H minus $r^2/(2^{r+k}-1)$, which tends to 0 as r tends to infinity. □

Returning to the proof of Proposition 1.1 we apply the lemma and take the union of the complement of the construction it provides in every (large) layer r of the hypercube. In the small layers we simply take all vertices. This gives a set of vertices of the n-cube that contains less than a fraction of 2^{-k} of the vertices and intersects every copy of $D_r(k+2\log k, 2k+3\log k)$. Since every subcube Q_d of Q_n of dimension $d = 2k+3\log k$ fully contains a copy of some $D_r(k+2\log k, 2k+3\log k)$ this completes the proof of the proposition. □

2.3 List Erasure Codes

In this subsection we describe the proofs of the improved upper and lower bounds for $L(d)$. The lower bounds follow easily from the results about graph and hypergraph augmentations proved in subsection 2.1. The upper bounds combine the construction in [4] with simple tools from linear algebra and a computation of the Lagrangians of appropriately defined t-uniform hypergraphs.

Starting with the proof of the lower bound define, for any integer $d \geq 2$, $g(d)$ to be the maximum possible value of the expression

$$\sum_{i=1}^{k+1} \prod_{j \in [k+1]-i} a_j$$

where the maximum is taken over all integers $k \geq 1$ and over all partitions of d of the form $d = a_1 + a_2 + \cdots + a_{k+1}$, where $a_i \geq 0$ are integers. Thus, for example, $g(2) = 2$ as demonstrated by the partition $2 = 2 + 0$, $g(5) = 8$ using the partition $5 = 2 + 2 + 1$ and $g(10) = 80$ using the partition $10 = 2 + 2 + 2 + 2 + 2$.

Lemma 2.2. *For every $d \geq 2$, $L(d) \geq g(d)$.*

Proof. Fix a small $\varepsilon > 0$ and let $\mathcal{C}$ be a collection of at least $\varepsilon \cdot 2^n$ vertices of Q_n. For a fixed $d \geq 2$ let $g(d) = \sum_{i=1}^{k+1} \prod_{j \in [k+1]-\{i\}} a_j$ where $a_j \geq 0$ are integers. Note that this number is exactly the number of edges of the k-uniform hypergraph H on $k+1$ vertex classes of sizes $a_1, a_2, \ldots, a_{k+1}$ whose edges are all k-tuples containing at most 1 vertex of each class. (This holds even if some of the numbers a_i are 0). By the remark following Propositions 1.4 and 1.5 if n is sufficiently large as a function of d and ε, then $\mathcal{C}$ must contain $H(r)$ for some r. The desired result follows as this $H(r)$ is fully contained in some subcube of dimension d in Q_n. Indeed, for $n \geq r + d - k$, fix a set of $d = a_1 + a_2 + \ldots + a_{k+1}$ coordinates partitioned into $k+1$ pairwise disjoint subsets $I_1, I_2, \ldots, I_{k+1}$ of cardinalities $a_1, a_2, \ldots, a_{k+1}$, respectively. Fix a binary vector x of Hamming weight $r - k$ on the other coordinates. The subcube Q_d in which the coordinates outside the set $\cup_i I_i$ are fixed to be x, and the coordinates in $\cup_i I_i$ are arbitrary, contains all the vectors of Hamming weight r that agree with x outside $\cup I_i$ and have exactly one 1-coordinate in each set I_i besides one of them. This is a copy of $H(r)$, as needed. □

The assertion of Proposition 1.2 follows easily from that of the last lemma. The bounds for $L(d)$ for $5 \leq d \leq 10$ are obtained by computing the value of $g(d)$ for these values of d. For large d divisible by 3, say $d = 3(k+1)$, it is not difficult to check that the value of $g(d)$ is obtained by the partition $a_1 = a_2 = \ldots = a_{k+1} = 3$, implying that $L(d) \geq (k+1)3^k = d \cdot 3^{(d-6)/3}$. □

We proceed with the proof of the upper bounds for $L(d)$ stated in Theorem 1.3, starting with several preliminary lemmas. For an integer $t \geq 1$, let $P(t)$ denote the probability that t binary vectors $v_1, v_2, \ldots, v_t$ in F_2^t, each chosen randomly, uniformly and independently among all $2^t - 1$ nonzero vectors in F_2^t, are linearly independent over F_2, that is, form a basis of F_2^t. Clearly $P(1) = 1$. Choosing the vectors one by one and multiplying the conditional probabilities that each vector is not spanned by the previously chosen ones assuming these are linearly independent, it follows that

$$P(t) = (\frac{2^t - 2}{2^t - 1}) \cdot (\frac{2^t - 4}{2^t - 1}) \cdots (\frac{2^t - 2^{t-1}}{2^t - 1}) = \frac{(2^t - 2)(2^t - 4) \cdots (2^t - 2^{t-1})}{(2^t - 1)^{t-1}}. \tag{1}$$

It is not difficult to check that for any $t > 1$

$$P(t) = (\frac{2^t - 2)}{2^t - 1})^{t-1} P(t-1). \tag{2}$$

This implies that for any k

$$c = \lim_{t \mapsto \infty} P(t) = \inf_t P(t) \geq P(k)(1 - O(\frac{k}{2^k})). \tag{3}$$

The equality (2) can be verified by induction on t, using (1). It can also be proved by the following combinatorial argument that will be useful later too.

The nonzero vectors $v_1, v_2, \ldots, v_t$ form a basis iff the following two events E_1 and E_2 hold. The event E_1 is that each v_i for $i \geq 2$ is not chosen to be equal to v_1. It is clear that its probability is exactly $(\frac{2^t-2}{2^t-1})^{t-1}$. Given the choice of v_1, each nonzero vector v in F_2^t has a unique expression as $v = x_v + y_v$, where $x_v \in \{0, v_1\}$ lies in the space generated by v_1, and y_v is orthogonal to this space. Let E_2 be the event that the vectors $y_{v_2}, y_{v_3}, \ldots, y_{v_t}$ form a basis of the $(t-1)$-dimensional subspace of F_2^t orthogonal to v_1. Conditioning on the event E_1, each nonzero vector of this $(t-1)$-dimensional space is selected with uniform probability among these $2^{t-1}-1$ possible vectors. These vectors span the space with probability $P(t-1)$, that is $Prob[E_2|E_1] = P(t-1)$. This implies (2) and hence also gives that the sequence $P(t)$ is monotone decreasing and thus approaches a limit, which is denoted by c in Theorem 1.3. It is easy to check that this limit is roughly 0.29.

We need the following simple result.

Lemma 2.3. *Let $t \geq 1$ and let $\{p_v : v \in F_2^t - \{0\}\ \}$ be an arbitrary probability distribution on the nonzero vectors in F_2^t. Then*

$$\sum_{v \in F_2^t - \{0\}} p_v (1-p_v)^{t-1} \leq (\frac{2^t-2}{2^t-1})^{t-1}.$$

Equality holds for the uniform distribution $p_v = 1/(2^t-1)$ for all $v \in F_2^t - \{0\}$.

Proof. The assertion is trivial for $t = 1$. For $t \geq 2$ put $g(z) = z(1-z)^{t-1}$. For $t = 2$ the second derivative of this function is $-2 < 0$ and hence it is concave in $[0,1]$, implying the desired result by Jensen's inequality. For $t \geq 3$ the derivative and second derivative of $g(z)$ are given by $g'(z) = (1-z)^{t-2}(1-tz)$ and $g''(z) = (1-z)^{t-3}(-2t+2+t(t-1)z)$. Therefore, in $[0,1]$ the function $g(z)$ is increasing in $[0,1/t)$, attains its maximum at $z = 1/t$, and is decreasing in $[1/t, 1]$. It is concave in $[0, 2/t)$ and convex in $[2/t, 1]$. Suppose that the sum $\sum_v g(p_v)$ considered in the lemma attains its maximum at $(p_v : v \in F_2^t - \{0\})$ (the maximum is clearly attained, by compactness). If there is some $p_v > 1/t$ then since $2^t - 1 > t$ there is also some $p_{v'} < 1/t$. Decreasing p_v by ε and increasing $p_{v'}$ by ε, for a sufficiently small $\varepsilon > 0$, strictly increases both $g(p_v)$ and $g(p_{v'})$, contradicting maximality. Therefore $0 \leq p_v \leq 1/t$ for all v. Since the function $g(z)$ is concave in $[0, 1/t]$ the maximum value of $\sum_v g(p_v)$ is obtained when all the values p_v are equal, by Jensen's Inequality. □

Corollary 2.4. *Let $t \geq 1$ and let $\{p_v : v \in F_2^t - \{0\}\ \}$ be an arbitrary probability distribution on the nonzero vectors in F_2^t. Then the probability that a sequence $v_1, v_2, \ldots, v_t$ of t random vectors, where each v_i is chosen randomly and independently according to this distribution, forms a basis of F_2^t is at most $P(t)$, where $P(t)$ is defined in (1). This is tight and obtained by the uniform distribution on $F_2^t - \{0\}$.*

Proof. We apply induction of t together with the reasoning described in the derivation of (2) from (1). The result is trivial for $t = 1$. Assuming it holds for $t-1$ we prove it for $t \geq 2$. Choosing the vectors $v_1, v_2, \ldots, v_t$ one by one, suppose $v_1 = v$ (this happens with probability p_v.) The vectors $v_1, v_2, \ldots, v_t$ form a basis iff no v_i for $i \geq 2$ is equal to $v_1 = v$ (denote this event by E_1), and the projections of the vectors $v_2, \ldots, v_t$ on the subspace orthogonal to v form a basis of this subspace (denote this event by E_2). The probability that $v_1 = v$ and E_1 holds is $p_v(1-p_v)^{t-1}$. The conditional probability that given this E_2 holds is, by the induction hypothesis, at most $P(t-1)$. Summing over v we conclude that the probability that $v_1, v_2, \ldots, v_t$ form a basis of F_2^t is at most

$$(\sum_{v \in F_2^v - \{0\}} p_v(1-p_v)^{t-1}) \cdot P(t-1).$$

The first factor is at most $(\frac{2^t-2}{2^t-1})^{t-1}$, by Lemma 2.3. This and (2) establish the desired inequality for t, completing the proof of the induction step and of the corollary. □

For integers $1 \leq k \leq d$ let $B(k,d)$ denote the maximum possible number of non-singular k by k submatrices in a k by d matrix over F_2. Therefore, $B(k,d)/\binom{d}{k}$ is the maximum possible probability that a set of k distinct columns of such a matrix forms a basis of F_2^k.

Lemma 2.5.

1. *For any fixed k the function $B(k,d)/\binom{d}{k}$ is monotone decreasing in d for all $d \geq k$, and is at least $P(k)$ for every admissible d.*

2. *For any $1 \leq k < d$, $B(k,d) = B(d-k,d)$.*

3. *For $d \geq k^2$*
$$(P(k) \leq) \frac{B(k,d)}{\binom{d}{k}} \leq \frac{P(k)d^k}{k!\binom{d}{k}} \leq P(k)(1 + \frac{k^2}{d-k}).$$

4. *For any $2 \leq k \leq d$,*
$$B(k,d) \leq \lfloor \frac{dB(k-1,d-1)}{k} \rfloor.$$

Proof.

1. Suppose $k \leq d' < d$. Let A be a k by d matrix over F_2 which maximizes the probability that a random set of k of its columns forms a basis. This probability is the average, over all choices of a k by d' submatrix A' of A, of the probability that a random set of k columns of A' forms a basis. The fact that
$$\frac{B(k,d')}{\binom{d'}{k}} \geq \frac{B(k,d)}{\binom{d}{k}}$$

follows by considering the submatrix A' maximizing this probability. To prove the inequality $B(k,d)/\binom{d}{k} \geq P(k)$ consider a random k by d matrix A over F_2 whose columns are chosen uniformly and independently in $F_2 - \{0\}$. Each subset of k columns of A is a basis with probability $P(k)$ and the desired inequality follows by linearity of expectation.

2. For a k by d matrix A of rank k over F_2, let A' denote the $(k-d)$ by d matrix whose rows form a basis of the subspace orthogonal to the row-space of A. If a set I' of $(d-k)$ columns of A' is of rank smaller than $d-k$ then there is a nonzero linear combination of the rows of A' which vanishes on these columns. This nonzero linear combination is orthogonal to the rows of A, providing a nontrivial linear relation of the columns $I = [d] - I'$ of A. This shows that if a set I' of $d-k$ columns of A' is not linearly independent, then the set $I = [d] - I'$ of k-columns of A is not linearly independent. By symmetry the converse holds as well, and the desired result follows by considering the matrices realizing $B(k,d)$ and $B(d-k,d)$.

3. Let A be a k by d matrix over F_2 with $B(k,d)$ nonsingular k by k submatrices. It is clear that A does not contain the 0-column (as it is easy to replace it and increase the number of nonsingular k by k submatrices). Let $\{p_v : v \in F_2^k - \{0\}\}$ be the probability distribution assigning to each column of A the same probability $1/d$. By Corollary 2.4 if we select k columns of A according to this probability distribution (with repetition), the probability we get a basis is at most $P(k)$. On the other hand this probability is exactly $k!(B(k,d)/d^k$. Therefore $\frac{k!B(k,d)}{d^k} \leq P(k)$ implying that

$$\frac{B(k,d)}{\binom{d}{k}} \leq \frac{P(k)d^k}{k!\binom{d}{k}} \leq P(k)e^{k(k-1)/2(d-k+1)} \leq P(k)[1 + \frac{k^2}{d-k}].$$

Here we used the fact that $\prod_{i=0}^{k-1}(d/d-i) < e^{k(k-1)/2(d-k+1)}$ and that $e^x \leq 1 + 2x$ for $x < 1$.

4. Let A be a k by d matrix with $B(k,d)$ k by k nonsingular submatrices. Every fixed column c of A can be contained in at most $B(k-1,d-1)$ such nonsigular matrices corresponding to the $(k-1)$ by $(k-1)$ nonsingular submatrices of the $(k-1)$ by $(d-1)$ matrix obtained from A by removing c and by replacing each column by its projection on the subspace orthogonal to c. The result thus follows by double counting.

□

Corollary 2.6. *Put $B(d) = \max\{B(k,d) : k \leq d$. Then $B(5) = 5$, $B(6) = 16$, $B(7) = 28$ and $B(d) = (c + o(1))\binom{d}{\lfloor d/2 \rfloor}$, where c is as in Theorem 1.3 and the $o(1)$-term tends to 0 as d tends to infinity.*

Proof. By Lemma 2.5, part 2, $B(d) = B(k, d)$ for some $k \leq d/2$. For $d = 5$ it is clear that $B(1, 5) \leq \binom{5}{1} = 5 < 8$. $B(2, 5)$ is the number of pairs of distinct columns of a 2 by 5 binary matrix in which every column is one of the three nonzero vectors of $F_2^2 - \{0\}$. This is clearly 8. The computation of $B(6) = B(3, 6)$ and of $B(7) = B(3, 7)$ is also simple and is obtained by any matrix with distinct columns in $F_2^3 - \{0\}$. (The upper bounds for these quantities also follow by Lemma 2.5, part 4 and the fact that $B(2, 4) = 5$).

To estimate $B(d)$ for large d observe, first, that by Lemma 2.5, part 1

$$B(d) \geq B(\lfloor d/2 \rfloor, d) \geq P(\lfloor d/2 \rfloor)\binom{d}{\lfloor d/2 \rfloor} > c\binom{d}{\lfloor d/2 \rfloor}.$$

Next, note that for, say $k < d/4$, $B(k, d) \leq \binom{d}{d/4}$ is (much) smaller than $c\binom{d}{\lfloor d/2 \rfloor}$, so $B(d) = B(k, d)$ for some $d/4 \leq k \leq d/2$. By Lemma 2.5, parts 1, 2 and 3, for any such k (and $d \geq k + \log k$):

$$\frac{B(k,d)}{\binom{d}{k}} \leq \frac{B(k, k+\log k)}{\binom{k+\log k}{k}} = \frac{B(\log k, k+\log k)}{\binom{k+\log k}{\log k}} \leq P(\log k)(1 + \frac{\log^2 k}{k})$$

$$\leq c(1 + O(\frac{\log k}{k}))(1 + \frac{\log^2 k}{k}) \leq c(1 + O(\frac{\log^2 d}{d})),$$

where in the penultimate inequality we used (3).

Therefore for each k in this range

$$B(k,d) \leq c\binom{d}{k}(1 + O(\frac{\log^2 d}{d})),$$

completing the proof. □

We are now ready to prove Theorem 1.3.

Proof. The relevant erasure codes are the ones constructed in [4], the novelty here is only in their analysis. Here is the description of the codes for a given length n. Let C_0 be the set consisting of the unique vector of weight 0 of Q_n. For each fixed r, $1 \leq r \leq n$, assign to each coordinate $i \in [n]$ a uniformly chosen random vector v_i in $F_2^r - \{0\}$. Let C_r denote the set of all binary vectors x of length n and Hamming weight r for which the r vectors v_i corresponding to all coordinates i with $x_i = 1$ form a basis of F_2^r. Note that the expected cardinality of C_r is $\binom{n}{r}P(r) > c\binom{n}{r}$, where c is the limit defined in Theorem 1.3 (which is roughly 0.29).

Fix a choice of vectors v_i for each r so that the resulting set C_r is of cardinality larger than $c\binom{n}{r}$ and let $\mathcal{C}$ be the union of all these sets. Thus $|\mathcal{C}| > c \cdot 2^n$.

Given $d < n$, partition $\mathcal{C}$ into $d + 1$ pairwise disjoint sets, where for each $0 \leq i \leq d$ the i-th sets consists of all vectors of $\mathcal{C}$ whose Hamming weight is i mod $(d + 1)$. Let $\mathcal{C}(d)$ denote the largest among those. Note that $|\mathcal{C}(d)| > (c/(d + 1))2^n$ contains a constant fraction of all binary vectors of length n. To complete the proof we prove an upper bound for the number of vectors of $\mathcal{C}(d)$ in any subcube of dimension d of Q_n.

Fix a subcube D of dimension d, and let $I \subset [n]$ be the set of the d coordinates that vary in the subcube. Observe that by the choice of the Hamming weights of the vectors in $\mathcal{C}(d)$, D can contain only vectors of $\mathcal{C}(d)$ of one specific Hamming weight. Denote this weight by r. Let x be the common projection of all points of D on $[n]-I$ and suppose its Hamming weight is $r-k$. Thus, each of the projections y of all the vectors of $\mathcal{C}(d)$ that lie in D on I has weight k. Let $v_1, v_2, \ldots, v_{r-k}$ be the binary vectors in $F_2^r - \{0\}$ that correspond to the indices i where $x_i = 1$. Then, by the construction of $C(r)$, these vectors are linearly independent. Moreover, for any vector y that appears as a projection above, the set of k vectors $v_j \in F_2^r - \{0\}$ that correspond to the coordinates j in which $y_j = 1$ complete the vectors $v_1, v_2, \ldots, v_{r-k}$ to a basis of F_2^r. This means that these vectors form a basis of the space of cosets of $W = \mathrm{span}(v_1, v_2, \ldots, v_{r-k})$ in F_2^r. This space is isomorphic to F_2^k. It follows that the number of such projections y is at most $B(k,d)$.

We have thus proved that the number of vectors of $\mathcal{C}(d)$ that lie in D is at most the maximum, over $k \leq d$, of the quantity $B(k,d)$, that is, at most $B(d)$. The desired upper bound thus follows from the last Corollary. □

2.4 Lagrangians

Some of the discussion in the previous subsection is equivalent to the computation of the Lagrangians of certain natural hypergraphs. Although this is not needed for the results here, we briefly describe the connection which may be of independent interest.

The Lagrangian Polynomial of a t-uniform hypergraph $H = (V,E)$ on a vertex set $V = \{1, 2, \ldots, n\}$ is the polynomial

$$P_H(x_1, x_2, \ldots, x_n) = \sum_{e \in E} \prod_{j \in e} x_j.$$

The Lagrangian $\lambda(H)$ of H is the maximum value of $P_H(x_1, \ldots, x_n)$ over the simplex $\{x_i \geq 0, \sum_i x_i = 1\}$ (this maximum is attained as the simplex is compact).

Lagrangians of hypergraphs were first considered by Frankl and Füredi [13] and by Sidorenko [16], extending the application of this notion for graphs, initiated by Motzkin and Straus [14].

For each $t \geq 1$ let B_t denote the t-uniform hypergraph on the vertex set $V = F_2^t - \{0\}$ of the $2^t - 1$ nonzero elements of the vector space of dimension t over F_2, whose edges are all bases of this vector space. Let $\lambda(B_t)$ denote the Lagrangian of this hypergraph. Trivially $\lambda(B_1) = 1$ and $\lambda(B_2) = 1/3$. By Corollary 2.4 for every fixed t, the value of the Lagrangian of B_t satisfies $\lambda(B_t) = P(t)/t!$. Therefore $\lambda(B_t) = (c + o(1))/t!$ where c is as in Theorem 1.3 and the $o(1)$-term tends to 0 as t tends to infinity.

3 Concluding remarks

- By Theorem 1.3 $L(5) = 8$. Therefore, for any arbitrarily small $\varepsilon > 0$, any set of at least an ε-fraction of the vertices of the n-cube for $n > n(\varepsilon)$ contains at least 8 vertices in some 5-dimensional subcube. On the other hand, there is a set of at least $c/6 > 0.04$-fraction of the vertices that does not contain more than 8 vertices of each such subcube. It is easy to improve this lower bound to $c/4 > 0.07$, since we can take the union of the subsets $C(r)$ for all Hamming weights r congruent to a constant modulo 4, instead of a constant modulo 6. It is easy to check that this still contains at most 8 vertices of any 5-subcube, since the sum of cardinalities of any two quantities $B(k, 5)$ for values of k that differ by at least 4 is at most 8. Similarly, for large d and any small $\varepsilon > 0$, the code containing all collections $C(r)$ for Hamming weights r congruent to a constant modulo $b(\varepsilon)\sqrt{d}$ has a fraction of $\Omega_\varepsilon(1/\sqrt{d})$ of all vertices of the cube and contains at most $(c + \varepsilon)\binom{d}{\lfloor d/2 \rfloor}$ vertices each d-subcube.

- As mentioned in the proof of Corollary 2.6 it is not difficult to find the exact values of $B(6) = 16$ and $B(7) = 28$. With a bit more work one can determine $B(d)$ precisely for larger (small) values of d, but since there is no reason to believe that these provide a tight bound for $L(d)$ we have not done that.

- The problem of determining the precise value of $L(d)$ for $d > 5$ remains open. It will be interesting to close the gap between the upper and lower bounds for these quantities. Another problem is the estimation of the largest possible cardinality of a $(d, L(d), n)$-code. As mentioned above, for $d = 5$ there is a $(5, 8, n)$-code containing more than $c/4 \cdot 2^n > 0.07 \cdot 2^n$ of the binary vectors of length n, but there is no reason to believe that this is tight. The analogous problem for $d = 4$, that is, determining the maximum possible fraction of the set of n-vectors in a $(4, 5, n)$-code, is also open. The lower bound here is $c/3 > 0.09$ and the trivial upper bound is $5/16$. For smaller values of d the analogous problem is not difficult. The even vectors are $1/2$ of the vectors, and form a $(1, 1, n)$-code and also a $(2, 2, n)$-code, and $1/2$ is clearly optimal here. For $d = 3$ the collection of all vectors of Hamming weight constant modulo 3 provide a $(3, 3, n)$-code with at least $1/3$ of all vectors. This $1/3$ is asymptotically optimal by the following argument. Let $\mathcal{C}$ be a collection of binary vectors of length n. If there are more than $2^{n-1}/n$ binary vectors v of length $n - 1$ so that both $v0$ and $v1$ are in $\mathcal{C}$, then there are two such vectors v, v' which differ by at most 2 coordinates, and in this case $\{v0, v1, v'0, v'1\}$ all lie in the same 3-cube, showing that $\mathcal{C}$ is not a $(3, 3, n)$-code. If not, and, say, $|\mathcal{C}|/2^n > (1/3 + 1/n)$ then the projection of $\mathcal{C}$ on the first $(n-1)$-coordinates is of cardinality exceeding $\lceil 2^n/3 \rceil$. By the result of [11] and [8] this projection contains a full 2-cube, implying that $\mathcal{C}$ contains at least 4 points in a 3-cube and showing it is not a $(3, 3, n)$-code.

Acknowledgment I would like to thank Matija Bucić, Maria-Romina Ivan, Robert Johnson and two anonymous reviewers for helpful comments.

References

[1] Noga Alon, Anja Krech and Tibor Szabó, Turán's theorem in the hypercube, SIAM J. on Disc. Math. 21 (2007), 66-72.

[2] Avraham Ben-Aroya, Dean Doron and Amnon Ta-Shma Near-Optimal Erasure List-Decodable Codes, Proc. CCC 2020, 1:1–1:27

[3] Béla Bollobás, Imre Leader and Claudia Malvenuto, Daisies and other Turán problems, Combin. Probab. Comput. 20 (2011), no. 5, 743–747.

[4] David Ellis, Maria-Romina Ivan and Imre Leader, Turań densities for daisies and hypercubes. arXiv:2401.16289, 2024.

[5] David Ellis and Dylan King, Lower bounds for the Turán densities of daisies, Electron. J. Combin. 30 (2023), no. 4, Paper No. 4.4, 11 pp.

[6] Venkatesan Guruswami, List decoding from erasures: Bounds and code constructions, IEEE Transactions on Information Theory, 49(11):2826-2833, 2003.

[7] Heiko Harborth and Hauke Nienborg, Some further vertex Turán numbers for cube graphs, Utilitas Math. 75 (2008), 83–87.

[8] Karen Anne Johnson and Roger Entringer, Largest induced subgraphs of the n-cube that contain no 4-cycles, J. Combin. Theory Ser. B 46 (1989), no. 3, 346–355.

[9] Robert Johnson, Private communication, 2024.

[10] Robert Johnson and John Talbot, Vertex Turán problems in the hypercube, J. Combin. Theory Ser. A 117 (2010), no. 4, 454-465.

[11] Alexandr Kostochka, Piercing the edges of the n-dimensional unit cube, (in Russian), Diskret. Analiz 1976, 55–64.

[12] Paul Erdős, On extremal problems of graphs and generalized graphs, Israel J. Math. 2 (1964), 183–190.

[13] Peter Frankl and Zoltán Füredi, Extremal problems and the Lagrange function for hypergraphs, Bull. Inst. Math. Acad. Sinica 16 (1988), no. 4, 305–313.

[14] Theodore S. Motzkin and Ernst Gabor Straus, Maxima for graphs and a new proof of a theorem of Turán, Canadian J. Math. 17 (1965), 533–540.

[15] Morris Plotkin, Binary codes with specified minimum distance, IRE Transactions on Information Theory 6 (4) (1960), 445–450.

[16] Alexander Sidorenko, On the maximal number of edges in a homogeneous hypergraph that does not contain prohibited subgraphs (Russian), Mat. Zametki 41 (1987), no. 3, 433–455.