Asymptotics of the Maximal Size of Almost Cancellative Families of Sets *

Vladimir B. Balakirsky

The author is currently with the
Fakultät für Mathematik, Universität Bielefeld,
Postfach 100131, D-33501 Bielefeld 1, Germany
on leave from the Data Security Association "Confident",
193060 St.-Petersburg, Russia
E-mail addresses :
vbal@mathematik.uni-bielefeld.de
vbal@vbal.spb.su

Abstract Asymptotic behavior of the maximal size of families $\mathcal{A}_n$ consisting of the sets $A, B, \ldots \subseteq \{1, \ldots, n\}$ and having the following property: B is uniquely determined by $(A, A \cup B)$ for at least $(1-\varepsilon)|\mathcal{A}_n|^2$ pairs $(A, B) \in \mathcal{A}_n \times \mathcal{A}_n$, is specified when $n \to \infty$ and $\varepsilon > 0$ is arbitrary small.

Key Phrases :
Cancellative families, Multi-user communication

*The work was partially supported by the SFB-343, Universität Bielefeld, Germany

1 Introduction and Formulation of the Results

Let $\mathcal{A}_n$ be a collection of subsets of the set $[n] = \{1, \ldots, n\}$.
Definition 1 [7] *$\mathcal{A}_n$ is a cancellative family if, for all $A, B \in \mathcal{A}_n$,*

$$A, A \cup B \ \ \textit{are known} \ \ \Longrightarrow \ \ B \ \ \textit{can be uniquely determined.} \tag{1}$$

Hence, if we are given a cancellative family then we have a possibility to "subtract" A from $A \cup B$ and obtain B. Note that, for any two sets A and B,

$$A \cup B = \overline{\overline{A} \cap \overline{B}},$$

where the line above some set denotes the complement to that set : $\overline{A} = [n] \backslash A$ etc. Therefore cancellative families can be also defined as follows : *$\mathcal{A}_n$ is a cancellative family if, for all $A, B \in \mathcal{A}_n$,*

$$\overline{A}, \overline{A} \cap \overline{B} \ \ \textit{are known} \ \ \Longrightarrow \ \ B \ \ \textit{can be uniquely determined.} \tag{2}$$

In other words, given a cancellative family $\mathcal{A}_n$, we can construct another family $\mathcal{A}'_n$ taking the complements of all sets of $\mathcal{A}_n$. If $A, B \in \mathcal{A}'_n$, then we have a possibility to "divide" $A \cap B$ by A and obtain B.

Let M_n denote the maximal number of subsets of a cancellative family constructed for a given n. The parameter

$$R = \lim_{n \to \infty} \frac{\log M_n}{n}$$

will be referred to as *asymptotic rate of cancellative families.* We will also denote by $M_{n,k}$ the maximal number of subsets of a cancellative family constructed for a given n when each subset has cardinality k.

A trivial construction of a cancellative family for $n = 3$ is

$$\mathcal{A}_{3,1} = \left\{ \ \{1\}, \{2\}, \{3\} \ \right\}, \tag{3}$$

where the index 1 shows that each subset consists of 1 element. This family can be concatenated : if n is divisible by 3 then for all $(i_1, \ldots, i_{n/3})$

$\in \{0,1,2\}^{n/3}$ we construct the set $\cup_{j=1}^{n/3}\{3j-2+i_j\}$ and include it into $\mathcal{A}_n$. Since $\mathcal{A}_n$ contains $3^{n/3}$ sets and has the cancellative property (1), we get a lower bound

$$R \geq \frac{\log 3}{3} = 0.528\ldots$$

Throughout the paper $\log z = \log_2 z$ for all $z > 0$. We also set $0 \log 0 = 0$.

If $n = 6$ then the result of a concatenation above is the family

$$\mathcal{A}_{6,2} = \Big\{ \ \{1,4\},\{1,5\},\{1,6\},\{2,4\},\{2,5\},\{2,6\},\{3,4\},\{3,5\},\{3,6\} \ \Big\}. \tag{4}$$

This family can be associated with the graph having n vertices that are numbered by integers $1,\ldots,n$ and any two vertices v_1 and v_2 are connected by an edge if and only if $\{v_1,v_2\} \in \mathcal{A}_{6,2}$ (see Fig.1). A similar graph can be introduced for any family $\mathcal{A}_{n,2}$. Then the cancellative property (1) implies that the graph does not contain cycles of length 3. An explicit construction of graphs having this property and maximal number of edges (complete bipartite graph) follows from Turán's theorem [9] and leads to cancellative families consisting of $\lfloor n^2/4 \rfloor$ subsets ($\lfloor z \rfloor$ denotes the maximal integer that is not greater than z). A generalization of the problem to the case when the subsets have cardinality $k > 2$ can be formulated in terms of hypergraphs. The solution

$$M_{n,k} = 2^{n-k}, \ \text{ for all } \ k \geq n/2, \tag{5}$$

and a construction of maximal families are known [7] : define $J(C) = \cup_{t=1}^{n-k}\{j_t(C)\}$ for all $C \subseteq [n-k]$, where $j_t(C) = t$ if $t \in C$ and $j_t(C) = t+n-k$ if $t \notin C$, and construct the family $\mathcal{A}_n$ consisting of subsets $[n]\backslash J(C)$, $C \subseteq [n-k]$. Another known explicit construction that can be presented as a construction of maximal cancellative families was derived by Bolobás [3]; it gives families consisting of $\lfloor n/3 \rfloor \lfloor (n+1)/3 \rfloor \lfloor (n+2)/3 \rfloor$ subsets of cardinality 3.

The family $\mathcal{A}_{3,1}$ can be also concatenated using an algorithm proposed by Shearer [8]. Let n be divisible by 3 and let $(i_1,\ldots,i_{n/3}) \in \{0,1,2,3\}^{n/3}$ be a vector satisfying the following restrictions : (a) there is exactly one component equal to 3; (b) the sum of all components is divisible by 3. Let us define the set $\cup_{j=1}^{n/3} D_j$, where $D_j = \{3j-2+i_j\}$ if $i_j \neq 3$ and $D_j = \{3j-2, 3j-1, 3j\}$

if $i_j = 3$, and include it into $\mathcal{A}_n$. Then taking all vectors $(i_1, \ldots, i_{n/3})$ satisfying the restrictions above we construct a cancellative family $\mathcal{A}_n$ consisting of $(n/3) \cdot 3^{n/3-2}$ sets. Thus,

$$R \geq \max_{l=2,3,\ldots} \frac{\log(l \cdot 3^{l-2})}{3l} = 0.547\ldots (l_0 = 24). \tag{6}$$

An upper bound

$$R \leq \max_{0.5 \leq \gamma \leq 1} [\ \gamma - 1 + h(\gamma)\ ] = \log\frac{3}{2} = 0.584\ldots (\gamma_0 = 2/3), \tag{7}$$

where

$$h(\gamma) = -\gamma \log \gamma - (1-\gamma)\log(1-\gamma)$$

is the binary entropy function, was proved by Frankl and Füredi [7]. Probably (6) and (7) are the best known bounds on the asymptotic rate R.

In our considerations we allow the error in determining B from $(A, A \cup B)$ and introduce the following

Definition 2 *For a given $\varepsilon > 0$, $\mathcal{A}_n$ is an ε-cancellative family if the fraction of subsets $A, B \in \mathcal{A}_n$, for which the cancellative property (1) holds, is at least $1 - \varepsilon$, i.e.,*

$$\lambda(\mathcal{A}_n) \leq \varepsilon, \tag{8}$$

where

$$\lambda(\mathcal{A}_n) = \frac{1}{M^2} \sum_{A \in \mathcal{A}_n} \sum_{B \in \mathcal{A}_n} \chi\left\{\ |\mathcal{C}(A \cup B|A)| > 1\ \right\}, \tag{9}$$

M denotes the number of subsets included into $\mathcal{A}_n$, and

$$\mathcal{C}(U|A) = \left\{\ B \in \mathcal{A}_n :\ A \cup B = U\ \right\},\ \ \textit{for all}\ \ U \supseteq A. \tag{10}$$

Hereafter χ stands for the indicator function : $\chi\{S\} = 1$ if the statement S is true and $\chi\{S\} = 0$ otherwise.

Let $M_n(\varepsilon)$ be the maximal number of subsets of an ε-cancellative family $\mathcal{A}_n$. The parameter

$$R' = \inf_{\varepsilon > 0} \lim_{n \to \infty} \frac{\log M_n(\varepsilon)}{n}$$

will be referred to as *asymptotic rate of almost cancellative families.* Hence, R' can be understood as the maximal rate of cancellative families such that the property (1) holds with arbitrary small error probability when the sets A and B are selected from $\mathcal{A}_n$ at random in accordance with the uniform distribution.

Theorem

$$R' = \max_{0\le\gamma\le 1} \gamma h_1(\gamma) = \max_{0\le\gamma\le 1} \gamma h(\gamma) = 0.616\ldots (\gamma_0 = 0.703\ldots),$$

where $h(\gamma)$ is the binary entropy function and

$$h_1(\gamma) = \begin{cases} 1, & \text{if } \gamma \in [0,1/2], \\ h(\gamma), & \text{if } \gamma \in (1/2,1]. \end{cases} \tag{11}$$

A curve of the function $\gamma h(\gamma), \gamma \in [0,1]$, is shown in Fig.2.

2 Basic Tools of the Proof and Relations to Multi-User Information Theory

One of transmission schemes analyzed in multi-user information theory is as follows. Suppose there are two users and one decoder. The t-th user has a binary $M_t \times n$ matrix (a code of length n and rate $R_t = (\log M)/n$) consisting of rows $\mathbf{x}_t(1), \ldots, \mathbf{x}_t(M_t) \in \{0,1\}^n$ and sends the vector $\mathbf{x}_t(m_t)$ to the OR multiple access channel if $m_t \in [M_t]$ is the message generated by the t-th source; $t = 1, 2$. The decoder receives the vector $\mathbf{x}_1(m_1) \vee \mathbf{x}_2(m_2)$ defined as a result of the component-wise OR operation and constructs an estimate $(\hat{m}_1, \hat{m}_2) \in [M_1] \times [M_2]$. The decoding error occurs if $(\hat{m}_1, \hat{m}_2) \neq (m_1, m_2)$, i.e., one of the following events takes place : (a) $\hat{m}_1 \neq m_1$ and $\hat{m}_2 = m_2$; (b) $\hat{m}_1 = m_1$ and $\hat{m}_2 \neq m_2$; (c) $\hat{m}_1 \neq m_1$ and $\hat{m}_2 \neq m_2$. The region consisting of all rate pairs (R_1, R_2) such that the decoding with arbitrary small error probability when the messages are random variables uniformly distributed on $[M_1]$ and $[M_2]$ was found by Ahlswede [1], [2] (see also [4], pp.388-403). Determining the maximal asymptotic rate of almost cancellative families R' can be viewed as determining the maximal rate for a multiple access channel above when 1) the codes of the users coincide; 2) the message of the second user is known to the decoder, i.e., only the error event (a) may take place. Therefore some parts of the proof of a general result for multiple access channels can be adopted for our purposes. This adaptation has some peculiarities coming with the problem and we consider them in this and the next section. The text is self-contained, and the proofs of auxiliary results are included.

Proof of the theorem is based on two statements. The direct statement claims that if each subset of $\mathcal{A}_n$ is chosen at random from the set

$$[n]_k = \Big\{ \ A \subseteq [n] : \ |A| = k \ \Big\}$$

in accordance with the uniform distribution, then the expectation of the error probability $\lambda(\mathcal{A}_n)$ tends to 0, as $n \to \infty$, if the rate $r' = (\log |\mathcal{A}_n|)/n$ is less than $\gamma h(\gamma)$, where

$$\gamma = \frac{n-k}{n}. \tag{12}$$

It means that there exists at least one sequence of families $\{\mathcal{A}_n\}$ with a fixed rate r' such that the corresponding sequence $\{\lambda(\mathcal{A}_n)\}$ also tends to 0, as $n \to \infty$. A converse statement claims that the rate $\gamma h_1(\gamma)$ is asymptotically the best possible.

Note that if $\gamma \in [0, 1/2]$, then there exist cancellative families, and hence almost cancellative families, having rate γ (see (5) and (12)). If we want to get this rate by our method of the analysis as an upper bound on the expected value of the rate of almost cancellative families chosen at random, then we should use a time sharing argument : construct a collection of families $\mathcal{A}_{2n\gamma,n\gamma}$ and add the elements $2n\gamma+1, \ldots, n$ to each subset. The converse statement claims that the rate γ cannot be increased in asymptotics, i.e., we do not gain by introducing the error in the cancellative property (1). However, if $\gamma \in (1/2, 1]$ it is the case, at least when $\gamma h(\gamma)$ is greater than the number at the right hand side of (7).

Let us consider essential steps of the proof of the direct statement and denote by $Q = (\, Q(A)\,)_{A \subseteq [n]}$ the uniform distribution on the set $[n]_k$, i.e.,

$$Q(A) = \begin{cases} \, |\, [n]_k \,|^{-1}, & \text{if } A \in [n]_k, \\ 0, & \text{if } A \notin [n]_k. \end{cases} \tag{13}$$

For a formal convenience, let us denote the entries of $\mathcal{A}_n$ by $A_1, \ldots, A_M$. Then we can rewrite (9), (10) as follows :

$$\lambda(\mathcal{A}_n) = \frac{1}{M^2} \sum_{m_0=1}^{M} \sum_{m=1}^{M} \chi \left\{ \exists m' \in [M] \backslash \{m\} : \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \right\}.$$

Note also that, for all $A_{m_0} \in [n]_k$,

$$\sum_{A_m \subseteq [n]} Q(A_m) \cdot \chi \left\{ \; |A_m \backslash A_{m_0}| = d \; \right\} \;\; = \;\; \binom{k}{d} \binom{n-k}{d} \binom{n}{k}^{-1}, \qquad d = 0, \ldots, \min\{k, n-k\}, \tag{14}$$

and if $|A_m \backslash A_{m_0}| = d$, then

$$\sum_{A_{m'} \subseteq [n]} Q(A_{m'}) \cdot \chi \left\{ \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \; \right\} \;\; = \;\; \binom{k}{d} \binom{n}{k}^{-1}. \tag{15}$$

Let $\mathbb{E}$ denote the averaging over all subsets $A_m, m \in [M]$, when each subset is independently chosen in accordance with probability distribution Q. Obviously, the main contribution to $\mathbb{E}\lambda(\mathcal{A}_n)$ give the terms included into the quantity

$$\mathbb{E}\lambda'(\mathcal{A}_n) = \mathbb{E}\frac{1}{M^2} \sum_{m_0=1}^{M} \sum_{\substack{m=1 \\ m \neq m_0}}^{M} \xi(\mathcal{A}_n | A_{m_0}, A_m), \tag{16}$$

where

$$\xi(\mathcal{A}_n | A_{m_0}, A_m) = \chi \left\{ \; \exists m' \in [M] \backslash \{m_0, m\} : \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \; \right\}, \tag{17}$$

i.e., we restrict our attention to the case when the error is caused by the subsets A_{m_0}, A_m, and $A_{m'}$ with three distinct indices. These subsets are independent random sets in our probabilistic ensemble.

The indicator function at the right hand side of (17) can be estimated by the union bound

$$\xi(\mathcal{A}_n | A_{m_0}, A_m) \leq \sum_{\substack{m'=1 \\ m' \neq m_0, m}}^{M} \chi \left\{ \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \; \right\}.$$

Substituting this bound in (16), we get

$$\mathbb{E}\lambda'(\mathcal{A}_n) \;\leq\; \frac{1}{M^2}\sum_{m_0=1}^{M}\sum_{\substack{m=1\\ m\neq m_0}}^{M}\sum_{A_{m_0}\subseteq[n]} Q(A_{m_0})\cdot\sum_{d=0}^{k}\sum_{A_m\subseteq[n]} Q(A_m)\cdot$$
$$\cdot\chi\Big\{\,|A_m\backslash A_{m_0}|=d\Big\}\cdot\mathbb{E}_{m_0,m}\xi(\mathcal{A}_n|A_{m_0},A_m),$$

where $\mathbb{E}_{m_0,m}$ denotes the averaging over all subsets $A_{m'}, m'\in[M]\backslash\{m_0,m\}$, when each subset is independently chosen in accordance with the probability distribution (13), and because of (14), (15),

$$\begin{aligned}&\mathbb{E}_{m_0,m}\xi(\mathcal{A}_n|A_{m_0},A_m)\\ &\quad\leq \mathbb{E}_{m_0,m}\sum_{\substack{m'=1\\ m'\neq m_0,m}}^{M}\chi\Big\{\;A_{m_0}\cup A_m = A_{m_0}\cup A_{m'}\;\Big\}\\ &\quad= (M-2)\binom{k}{d}\binom{n}{k}^{-1}\end{aligned}\tag{18}$$

and

$$\mathbb{E}\lambda'(\mathcal{A}_n)\leq (M-2)\sum_{d=0}^{k}\binom{n-k}{d}\left[\binom{k}{d}\binom{n}{k}^{-1}\right]^2. \tag{19}$$

Let $\alpha = d/n$ and let γ be defined in (12). Then taking the product of the maximal term in the sum at the right hand side of (19) and using the well-known estimates that are based on the Striling's approximation of the factorial (see, for example [6], p.530), we write

$$\binom{n-k}{d}\leq 2^{n\gamma h(\alpha/\gamma)},\quad \binom{k}{d}\leq 2^{n(1-\gamma)h(\alpha/(1-\gamma))},$$

and

$$\binom{n}{k}\geq\left(\frac{2}{\pi n}\right)^{1/2}2^{nh(\gamma)}.$$

Hence, we conclude that

$$-\frac{1}{n}\log\mathbb{E}\lambda'(\mathcal{A}_n)\geq E(1)-\frac{\log M}{n}-\delta_n,$$

where

$$E(1) = 2h(\gamma) - \max_{0 \le \alpha \le \gamma, 1-\gamma} \left[\gamma h\left(\frac{\alpha}{\gamma}\right) + 2(1-\gamma) h\left(\frac{\alpha}{1-\gamma}\right)\right] \tag{20}$$

and

$$\delta_n = \frac{1}{n} \log \frac{\pi n(n+1)}{4} \to 0, \quad \text{as } n \to \infty. \tag{21}$$

The maximum on α at the right hand side of (20) is attained when $\alpha = \alpha_0 < \gamma(1-\gamma)$ and these considerations do not give desired asymptotic rate since $E(1) < \gamma h(\gamma)$. Thus, we need more tight estimates and use the approach developed by Gallager for the proof of a random coding bound for memoryless channels [5] (see also [6], pp.135-137).

Proposition 1 *If $A_{m_0}, A_m \subseteq [n]_k$ are fixed in such a way that $|A_m \backslash A_{m_0}| = d$, then*

$$\mathbb{E}_{m_0,m} \xi(\mathcal{A}_n | A_{m_0}, A_m) \quad \le \quad \min_{\rho \in [0,1]} (M-2)^\rho \left[\binom{k}{d} \binom{n}{k}^{-1} \right]^\rho . \tag{22}$$

Proof Obviously,

$$\xi(\mathcal{A}_n | A_{m_0}, A_m) \quad \le \quad \left[\sum_{\substack{m'=1 \\ m' \ne m_0, m}}^{M} \chi \left\{ \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \; \right\} \right]^\rho ,$$

for all $\rho \ge 0$.

If $\rho \in [0,1]$, then using Jensen's inequality ([6], p.523) we obtain

$$\mathbb{E}_{m_0,m} \xi(\mathcal{A}_n | A_{m_0}, A_m) \quad \le \quad \left[\mathbb{E}_{m_0,m} \sum_{\substack{m'=1 \\ m' \ne m_0, m}}^{M} \chi \left\{ \; A_{m_0} \cup A_m = A_{m_0} \cup A_{m'} \; \right\} \right]^\rho$$

and using (18) complete the proof.

□

Repeating the previous considerations with the upper bound (22) instead of (18) we obtain

$$-\frac{1}{n} \log \mathbb{E} \lambda'(\mathcal{A}_n) \ge E(\rho) - \rho \frac{\log M}{n} - \delta_n, \tag{23}$$

where

$$E(\rho) = (1+\rho)h(\gamma) - \max_{0 \leq \alpha \leq \gamma, 1-\gamma} \left[\gamma h\left(\frac{\alpha}{\gamma}\right) + (1+\rho)(1-\gamma)h\left(\frac{\alpha}{1-\gamma}\right)\right]$$

and δ_n is defined in (21). Now we note that

$$\rho \to 0 \implies \alpha_0 \to \gamma(1-\gamma), \quad E(\rho) \to \rho\gamma h(\gamma). \tag{24}$$

Hence, if $(\log M)/n$ is a constant and $\gamma h(\gamma) - (\log M)/n > 0$, then we can assign ρ as a function of n in such a way that

$$\rho \to 0, \quad n\left(E(\rho) - \rho\frac{\log M}{n} - \delta_n\right) \to \infty, \qquad \text{as } n \to \infty.$$

Formulation of the statement that follows from these considerations and analysis of the remaining terms of $\lambda(\mathcal{A}_n)$ are presented in the next section.

Let us consider a converse statement. Note that $\gamma h(\gamma)$ coincides with the conditional entropy

$$H(Z|X) = -\sum_{x,z \in \{0,1\}^2} \pi(x)\omega(z|x)\log\omega(z|x)$$

constructed in the ensemble defined by the probability distribution

$$\Big(\ \pi(x)\omega(z|x)\ \Big)_{(x,z)\in\{0,1\}^2}, \tag{25}$$

where $(\pi(0), \pi(1)) = (\gamma, 1-\gamma)$ and

$$\omega(z|x) = \sum_{y \in \{0,1\}} \pi(y)\chi\{\ z = x \vee y\ \} = \begin{cases} \gamma, & \text{if } (x,z) = (0,0), \\ 1-\gamma, & \text{if } (x,z) = (0,1), \\ 0, & \text{if } (x,z) = (1,0), \\ 1, & \text{if } (x,z) = (1,1). \end{cases}$$

Note also that $\gamma h_1(\gamma)$ is a convex up closure of $\gamma h(\gamma)$.

Analysis of the number of subsets M of an ε-cancellative families in the next section uses some simple transformations and the well-known upper bound [6], pp.75-76, presented below.

Proposition 2 *Let*

$$P = (\, P(X)\,)_{X \subseteq [n]}$$

be a probability distribution on the set $[n]$ *and let* $(1-p_t, p_t)$ *be a projection of* P *on the element* $\{t\}$*, i.e.,*

$$p_t = \sum_{X \subseteq [n]} P(X) \cdot \chi\{\, t \in X \,\}, \ \ t = 1, \ldots, n.$$

Then

$$-\sum_{X \subseteq [n]} P(X) \cdot \log P(X) \ \ \leq \ \ \sum_{t=1}^{n} h(p_t),$$

where h *is the binary entropy function.*

Proof Let us introduce a probability distribution

$$\hat{P} = \left(\ \hat{P}(X) = \prod_{t=1}^{n} p_t(X) \right)_{X \subseteq [n]},$$

where

$$p_t(X) = \begin{cases} 1 - p_t, & \text{if } \ t \notin X, \\ p_t, & \text{if } \ t \in X. \end{cases}$$

Then using the inequality $\log z \leq (z-1)\ln 2$, $z > 0$, we have

$$\begin{aligned} -\sum_{X \subseteq [n]} P(X) \cdot \log P(X) \ &= \ \sum_{X \subseteq [n]} P(X) \cdot \left[\log \frac{\hat{P}(X)}{P(X)} - \log \hat{P}(X) \right] \\ &\leq \ -\sum_{X \subseteq [n]} P(X) \cdot \log \hat{P}(X) \\ &= \ -\sum_{t=1}^{n} \sum_{X \subseteq [n]} P(X) \cdot \log p_t(X) \\ &= \ -\sum_{t=1}^{n} p_t(X) \cdot \log p_t(X) \\ &= \ \sum_{t=1}^{n} h(p_t). \end{aligned}$$

$\square$

As a result, we estimate M via the conditional entropies $H(Z_1|X_1), \ldots,$ $H(Z_n|X_n)$ for n independent pairs of random variables (X_t, Z_t) defined by the probability distributions

$$\Big(\; \pi_t(x)\omega_t(z|x) \; \Big)_{(x,z)\in\{0,1\}^2},$$

where $(\pi_t(0), \pi_t(1)) = (\gamma_t, 1-\gamma_t)$ and

$$\omega_t(z|x) = \sum_{y\in\{0,1\}} \pi_t(y)\chi\{\; z = x \vee y \;\}, \;\; t = 1, \ldots, n.$$

However, the conditional distribution $(\,\omega_t(z|x)\,)_{(x,z)\in\{0,1\}^2}$ depends on the input distribution $(\,\pi_t(x)\,)_{x\in\{0,1\}}$ and we cannot upper-bound the sum $H(Z_1|X_1)$ $+\ldots+H(Z_n|X_n)$ by the product of n and the conditional entropy $H(Z|X)$ defined in the ensemble (25) with $\pi(x) = (\pi_1(x)+\ldots+\pi_n(x))/n$, $x \in \{0,1\}$. It is possible only after convexification of the conditional entropy function.

3 Proof of the Theorem

Direct statement : *Let $\varepsilon, \Delta > 0$ be fixed and let γ be a rational number from the interval $[0,1]$. If*

$$r' = \gamma h(\gamma) - \Delta, \tag{26}$$

then one can find an $n_0(\varepsilon,\Delta) < \infty$ such that, for all $n \geq n_0(\varepsilon,\Delta)$, $n\gamma$ is an integer, there exists an ε-cancellative family $\mathcal{A}_n^*$ consisting of $2^{nr'}$ subsets of the set $[n]$ having cardinality $n\gamma$.

Proof Let each subset of $\mathcal{A}_n$ be an independent random set chosen from $[n]_k$ in accordance with the uniform distribution, where $k = n(1-\gamma)$. Then, as it is easy to see,

$$\mathbb{E}\lambda(\mathcal{A}_n) \leq \mathbb{E}\lambda'(\mathcal{A}_n) + \mathbb{E}\lambda''(\mathcal{A}_n),$$

where $\mathbb{E}\lambda'(\mathcal{A}_n)$ is defined in (16) and

$$\begin{aligned} \mathbb{E}\lambda''(\mathcal{A}_n) \;\; &= \;\; \mathbb{E}\frac{1}{M^2}\sum_{m_0=1}^{M} \chi\Big\{\; \exists m' \in [M]\backslash\{m\} : \; A_{m_0} = A_{m_0} \cup A_m \Big\} \\ &\quad + \mathbb{E}\frac{1}{M^2}\sum_{m_0=1}^{M}\sum_{\substack{m=1\\ m\neq m_0}}^{M} \chi\Big\{\; A_{m_0} \cup A_m = A_{m_0} \Big\} \end{aligned}$$

$$\leq \; 2\mathbb{E}\frac{1}{M^2}\sum_{m_0=1}^{M}\sum_{\substack{m=1\\ m\neq m_0}}^{M}\chi\Big\{\, A_{m_0}\cup A_m = A_{m_0}\Big\}$$

$$= \; 2\cdot\binom{n}{k}^{-1},$$

i.e., $\mathbb{E}\lambda''(\mathcal{A}_n)$ can be upper-bounded by an exponentially decreasing function of n. Combining this bound with (23) and (24) we obtain that if the condition (26) is satisfied, then one can find an $n_0(\varepsilon,\Delta)<\infty$ such that

$$\mathbb{E}\lambda(\mathcal{A}_n)\leq\varepsilon, \;\text{ for all }\; n\geq n_0(\varepsilon,\Delta),\; n\gamma \text{ is an integer.}$$

However,

$$\mathbb{E}\lambda(\mathcal{A}_n)\leq\varepsilon \;\Longrightarrow\; \exists\mathcal{A}_n^*:\; \lambda(\mathcal{A}_n^*)\leq\varepsilon.$$

$\square$

Converse statement : *Let $\mathcal{A}_n$ be an ε-cancellative family consisting of M subsets of the set $[n]$. Then*

$$\frac{\log M}{n}\leq\frac{1}{1-\varepsilon}\cdot\max_{0\leq\gamma\leq 1}\gamma h_1(\gamma), \tag{27}$$

where the function h_1 is defined in (11).

Proof Using (8), (9) we write

$$\frac{1}{M^2}\sum_{A\in\mathcal{A}_n}\sum_{B\in\mathcal{A}_n}\log\frac{|\mathcal{C}(A\cup B|A)|}{M}$$

$$=\frac{1}{M^2}\sum_{A\in\mathcal{A}_n}\sum_{B\in\mathcal{A}_n}\sum_{U\supseteq A}\chi\{\,A\cup B=U\,\}\cdot\log\frac{|\mathcal{C}(U|A)|}{M}$$

$$=\frac{1}{M}\sum_{A\in\mathcal{A}_n}\sum_{U\supseteq A}\left(\frac{1}{M}\sum_{B\in\mathcal{A}_n}\chi\{\,A\cup B=U\,\}\right)\cdot\log\frac{|\mathcal{C}(U|A)|}{M}$$

$$=\frac{1}{M}\sum_{A\in\mathcal{A}_n}\sum_{U\supseteq A}\frac{|\mathcal{C}(U|A)|}{M}\cdot\log\frac{|\mathcal{C}(U|A)|}{M},$$

$$
\begin{aligned}
&\frac{1}{M^2}\sum_{A\in\mathcal{A}_n}\sum_{B\in\mathcal{A}_n}\log|\mathcal{C}(A\cup B|A)| \\
&\qquad\qquad \le \frac{1}{M^2}\sum_{A\in\mathcal{A}_n}\sum_{B\in\mathcal{A}_n}\chi\{\ |\mathcal{C}(A\cup B|A)|>1\ \}\cdot\log M \\
&\qquad\qquad \le \varepsilon\cdot\log M.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\log M &= \frac{1}{M^2}\sum_{A\in\mathcal{A}_n}\sum_{B\in\mathcal{A}_n}\log\left(\frac{M}{|\mathcal{C}(A\cup B|A)|}\cdot|\mathcal{C}(A\cup B|A)|\right) \qquad (28)\\
&\le -\frac{1}{M}\sum_{A\in\mathcal{A}_n}\sum_{U\supseteq A}\frac{|\mathcal{C}(U|A)|}{M}\cdot\log\frac{|\mathcal{C}(U|A)|}{M}+\varepsilon\cdot\log M.
\end{aligned}
$$

Note that $P_A = \Big(\ |\mathcal{C}(U|A)|/M\Big)_{U\supseteq A}$ is a conditional probability distribution on the sets containing A since every set $B\in\mathcal{A}_n$ belongs to exactly one collection $\mathcal{C}(U|A)$ (namely, $\mathcal{C}(A\cup B|A)$) and

$$
\sum_{U\supseteq[n]}|\mathcal{C}(U|A)| = M.
$$

Thus, using Proposition 2 we write

$$
-\frac{1}{M}\sum_{U\supseteq A}\frac{|\mathcal{C}(U|A)|}{M}\cdot\log\frac{|\mathcal{C}(U|A)|}{M} \quad\le\quad \sum_{t=1}^{n}h(p_{t,A}), \qquad (29)
$$

where

$$
\begin{aligned}
1-p_{t,A} &= \sum_{U\supseteq A}\frac{|\mathcal{C}(U|A)|}{M}\cdot\chi\{\ t\notin U\ \}. \qquad (30)\\
&= \frac{1}{M}\sum_{U\supseteq A}\sum_{B\in\mathcal{A}_n}\chi\{\ A\cup B=U\ \}\cdot\chi\{\ t\notin U\ \}\\
&= \frac{1}{M}\sum_{B\in\mathcal{A}_n}\chi\{\ t\notin A\cup B\ \}.\\
&= \chi\{\ t\notin A\ \}\cdot\gamma_{t,\mathcal{A}_n}
\end{aligned}
$$

and

$$
\gamma_{t,\mathcal{A}_n} = \frac{1}{M}\sum_{B\in\mathcal{A}_n}\chi\{\ t\notin B\ \} \qquad (31)
$$

does not depend on A.

Based on (29)-(31) and the equation $h(z) = h(1-z), z \in [0,1]$, we continue (28) as follows :

$$\begin{aligned}(1-\varepsilon)\log M &\leq \frac{1}{M}\sum_{A\in\mathcal{A}_n}\sum_{t=1}^{n}\chi\{\, t \notin A \,\}\cdot h(\gamma_{t,\mathcal{A}_n}) \\ &= \sum_{t=1}^{n}\left(\frac{1}{M}\sum_{A\in\mathcal{A}_n}\chi\{\, t\notin A\,\}\right)\cdot h(\gamma_{t,\mathcal{A}_n}) \\ &= \sum_{t=1}^{n}\gamma_{t,\mathcal{A}_n}h(\gamma_{t,\mathcal{A}_n}) \\ &\leq \sum_{t=1}^{n}\gamma_{t,\mathcal{A}_n}h_1(\gamma_{t,\mathcal{A}_n}),\end{aligned} \tag{32}$$

where the function h_1 is defined in (11) and the last inequality is introduced because $zh(z)$ is not a convex up function of z for all $z \in [0,1]$. However, this is true for the function $zh_1(z)$ and we have

$$\sum_{t=1}^{n}\gamma_{t,\mathcal{A}_n}h_1(\gamma_{t,\mathcal{A}_n}) \leq n\gamma_{\mathcal{A}_n}h_1(\gamma_{\mathcal{A}_n}), \tag{33}$$

where

$$\gamma_{\mathcal{A}_n} = \frac{1}{n}\sum_{t=1}^{n}\gamma_{t,\mathcal{A}_n}.$$

Since

$$\gamma_{\mathcal{A}_n}h_1(\gamma_{\mathcal{A}_n}) \leq \max_{0\leq\gamma\leq 1}\gamma h_1(\gamma),$$

(32) and (33) imply (27).

□

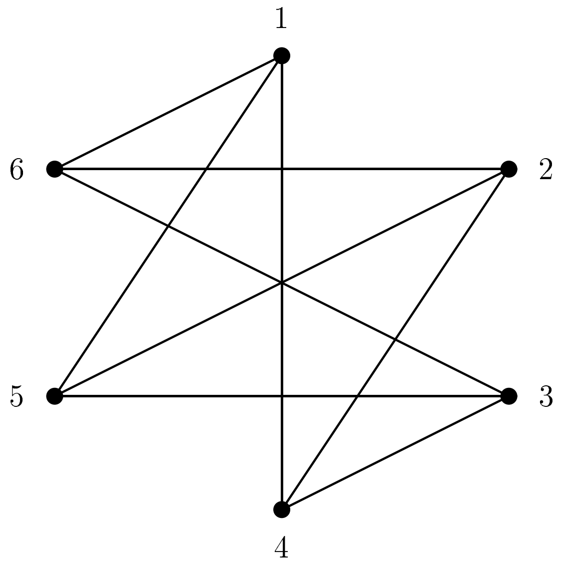

Figure 1: Graph corresponding to the family $\mathcal{A}_{6,2}$ defined in (4) (cycles of length 3 are absent).

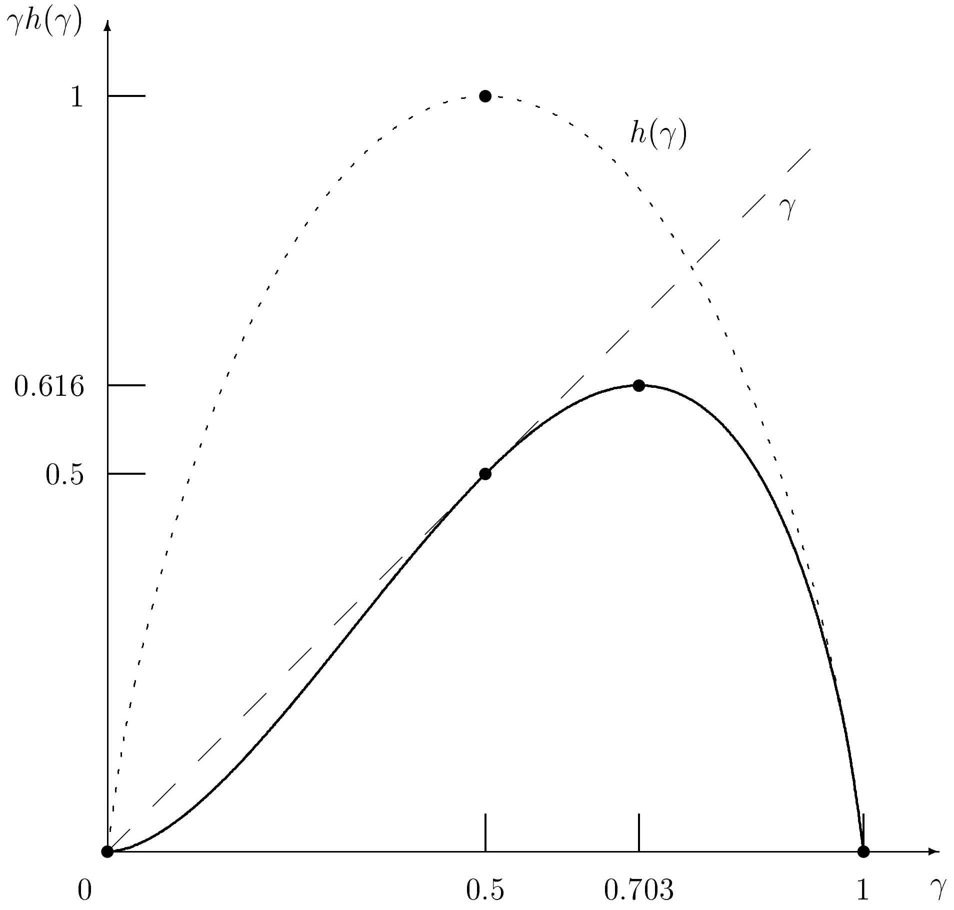

Figure 2: Curve of the function $\gamma h(\gamma)$.

References

[1] R.Ahlswede, Multi-way communication channels. *Proceedings of the 2nd International Symposium on Information Theory,* Tsahkadzor, Armenian SSR, 1971. Publishing House of the Hungarian Academy of Sciences, 23-52 (1973).

[2] R.Ahlswede, The capacity region of a channel with two senders and two receivers, *Ann.Prob.,* 2(5), 805-814 (1974).

[3] B.Bolobás, Three-graphs without two triples whose symmetric difference is contained in a third, *Discr.Math.,* 8, 21-24 (1974).

[4] T.M.Cover and J.A.Thomas, *Elements of Information Theory,* Wiley, New York, 1991.

[5] R.G.Gallager, A simple derivation of the coding theorem and some applications, *IEEE Trans.Inform.Theory,* 11(1), 3-18 (1965).

[6] R.G.Gallager, *Information Theory and Reliable Communication,* Wiley, New York, 1968.

[7] P.Frankl and Z.Füredi, Union-free hypergraphs and probability theory, *European Journal of Combinatorics,* 5, 127-131 (1984).

[8] J.B.Shearer, A new construction for cancellative families of sets, *The Electronic Journal of Combinatorics,* 3 (1996).

[9] P.Turán, On the theory of graphs, *Colloq.Math.,* 3, 19-30 (1954).