Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Discrete Applied Mathematics 128 (2003) 11–26

DISCRETE APPLIED MATHEMATICS

www.elsevier.com/locate/dam

Application of cover-free codes and combinatorial designs to two-stage testing

Toby Berger[a,1], Vladimir I. Levenshtein[b,*,2]

[a] *School of Electrical Engineering, Cornell University, Ithaca, NY 14853, USA*
[b] *Keldysh Institute for Applied Mathematics, Russian Academy of Sciences, Miusskaya sq. 4, 125047 Moscow, Russia*

Received 23 February 2001; received in revised form 22 March 2002; accepted 8 April 2002

Abstract

We study combinatorial and probabilistic properties of cover-free codes and block designs which are useful for their efficient application as the first stage of two-stage group testing procedures. Particular attention is paid to these procedures because of their importance in such applications as monoclonal antibody generation and cDNA library screening.

Keywords: Block designs; Cover-free codes; Steiner systems; Two-stage testing

1. Introduction

We consider the theory and design of efficient combinatorial and probabilistic group testing procedures for a population of size v. Each item in the population either is active or is inactive. When a subset of such items is tested as a group, the test's outcome will be 1 if at least one of the items in the group is active and 0 otherwise. Each such group test can be represented as a row in a v-column matrix, with a 1 in each column indexed by an item that is in the group and a 0 in the remaining columns. Simultaneously conducting b such group tests constitutes a *stage* of testing and is represented by a $b \times v$ matrix, A, with entries from $\{0, 1\}$. With a view toward

[*] Corresponding author.
E-mail address: leven@keldysh.ru (V.I. Levenshtein).

[1] Supported by NSF Grants ECS-9632266 and CCR-9980616.

[2] The research was partially supported by the Russian Foundation for Basic Research under grant 01-01-00035.

PII: S0166-218X(02)00433-X

strengthening the connection between group testing and error control codes, we define the *syndrome* of A to be the b-dimensional binary column vector Y that results from the disjunctive multiplication of A and the unknown v-dimensional binary column vector X whose ith component is 1 if item i is active and 0 if it is inactive. The problem is to determine X from analysis of Y. Often, some components of X remain undetermined (unresolved), in which case one or more subsequent stages of testing are needed. A subsequent stage can be modeled by adjoining one or more rows to A and thereby augmenting the syndrome.

When X has probability distribution P, we denote by $E(A,P)$ the expected number of tests in a two-stage model that uses A for the first stage and whose second stage consists of simultaneous individual tests of each item not resolved by stage 1. A sequence of A-matrices indexed by increasing v is called *asymptotically good* if $E(A,P)/v \to 0$ as $v \to \infty$. We introduce a new parameter $t^{+}(A)$, the maximal number t such that one can determine solely from the syndrome of A whether the number of active items exceeds t or not. We prove that $t^{+}(A)$ equals the maximal t such that the columns of A form a t-cover-free code. Then we exhibit sequences of 2-stage tests based on t-cover-free codes that are asymptotically good for certain sequences of probability measures governing $X \in \{0,1\}^{v}$. Some of our results are: (i) there are no asymptotically good matrices for the Bernoulli scheme with a constant $p, 0 < p < 1$, a result also obtainable via information-theoretic reasoning, (ii) there exist asymptotically good sequences based on cover-free codes for the Bernoulli p-scheme when $p = p(v) = o(1/\sqrt{v \ln v})$. (iii) The condition $c < \frac{1}{4}$ is necessary and sufficient for the Bernoulli p-scheme with $p(v) = (c \ln v)/\sqrt{v}$ to be among those situations in which there are asymptotically good matrices that are (v,k,b,r)-designs.

2. Classification of item states by a syndrome

We consider a set N_v comprised of v elements; henceforth, we refer to the elements as *items*. Without loss of generality let $N_v = \{1,2,\dots,v\}$. For any subset $X \subseteq N_v$ we denote by $\mathbf{X}$ the indicator vector $(x_1,\dots,x_v)^{\mathrm{T}}$, where $x_j = 1$ if $j \in X$, and $x_j = 0$ otherwise. We call the item j *active* if $x_j = 1$ and *inactive* if $x_j = 0$. To ascertain, or reconstruct, an unknown $X \subseteq N_v$, we shall use a set $A_1,\dots,A_b$ of non-empty subsets of N_v which are called *pools* or *group tests*. Each pool A_i also can be described by a binary vector $\mathbf{A}_i = (a_{i,1},\dots,a_{i,v})$, where $a_{i,j} = 1$ if $j \in A_i$, and $a_{i,j} = 0$ if otherwise. We denote by A the matrix $(a_{i,j})$ of size $b \times v$ and assume it has no all-zeros rows. Given $X \subseteq N_v$, we call pool A_i and the corresponding test *negative* and write $y_i = 0$ if $A_i \cap X$ is empty; if A_i is not negative, we call pool A_i and the corresponding test *positive* and write $y_i = 1$. The column vector $\mathbf{Y} = (y_1,\dots,y_b)^{\mathrm{T}}$ will be referred to as a *syndrome* by analogy with the theory of binary linear error-correcting codes. We denote by Y the subset of N_b consisting of the numbers of the unit coordinates of $\mathbf{Y}$. Since

$$y_i = a_{i,1}x_1 \vee \cdots \vee a_{i,v}x_v,$$

we may also write

$$\mathbf{Y} = A\mathbf{X} \tag{1}$$

with the understanding that disjunction is used in the matrix—vector product $A\mathbf{X}$ instead of modulo two sum. Thus, if $X = \{j_1,\dots,j_t\}$, then the syndrome $\mathbf{Y} = A\mathbf{X}$ is the componentwise disjunction of t columns $\mathbf{B}_{j_h} = (a_{1,j_h},\dots,a_{b,j_h})$ of $A, h = 1,\dots,t$. (We shall denote by B_j the subset of N_b consisting of the coordinate positions of the unit entries of the column $\mathbf{B}_j$.)

The setting of our reconstruction problem depends on whether X is chosen from the set $Q(v)$ of all 2^v subsets of N_v or only from a certain subset $Q \subset Q(v)$; examples are the subset $Q_t^-(v)$ consisting of all subsets of N_v with t or fewer elements, and the subset $Q_t^+(v)$ consisting of all subsets of N_v with t or more elements. For a matrix A and a syndrome $\mathbf{Y}$, denote by $Q(A,\mathbf{Y})$ the (possibly empty) set of all $X \subseteq N_v$ such that (1) holds.

Given $Q \subseteq Q(v)$, we say that the pools $\{A_i,\ i = 1,\dots,b\}$ solve the reconstruction problem for the set Q, or equivalently that the corresponding matrix A solves it, if any two distinct members of Q have different syndromes. If a matrix A solves this problem for the set $Q_t^-(v)$, then the set of its columns is called a *disjunctive* (v,b,t)-code. For any matrix A (without zero columns) denote by $t^-(A)$ the maximum number t such that A solves the reconstruction problem for $Q_t^-(v)$, or equivalently, knowledge of the syndrome $\mathbf{Y} = A\mathbf{X}$ (or the corresponding set Y) allows us to ascertain the set $X \subseteq N_v$ if $|X| \leqslant t$. Since syndromes are binary vectors, the number b of pools constituting a solution of the problem for a set $Q \subseteq Q(v)$ must satisfy the inequality

$$b \geqslant \log_2 |Q|. \tag{2}$$

This bound is attained when Q is the set $Q(v)$ of all 2^v subsets of N_v, since one can test each of the N_v items individually, i.e., set A equal to the v-dimensional identity matrix up to a permutation of rows. However, we shall see that this bound is not good in general. In particular, we shall verify that the set $Q_{v-1}^+(v)$ of cardinality $v+1$ requires use of v pools whereas (2) gives only $b \geqslant \log_2(v+1)$ in this case.

If the matrix A does not solve the problem for the set $Q \subseteq Q(v)$, we can consider it as the first stage of an *adaptive* testing algorithm. An adaptive algorithm recursively chooses a certain collection of new pools that depends on all previous pools and their syndromes. The maximum over any $X \in Q$ of the number of choices of collections of new pools needed to ensure correct reconstruction of X is called the *number of stages* of the adaptive algorithm. In particular, an adaptive algorithm might use only one additional test at each stage.

We now introduce another important characteristic of a matrix A (without zero columns). Denote by $t^+(A)$ the maximum integer t such that for any $\mathbf{Y}$,

$$Q(A,\mathbf{Y}) \cap Q_t^-(v) = \emptyset \quad \text{or} \quad Q(A,\mathbf{Y}) \cap Q_{t+1}^+(v) = \emptyset \tag{3}$$

(here $\emptyset$ is the empty set). Such a number $t^+(A)$ exists, since Eq. (3) holds for $t = 0$. Moreover, it is clear that, for any A and $\mathbf{Y}$, (3) holds for any t in the range $0 < t < t^+(A)$ (otherwise, if $X_1 \in Q(A,\mathbf{Y}), |X_1| \leqslant t,\ X_2 \in Q(A,\mathbf{Y}), t+1 \leqslant |X_2| \leqslant t^+(A)$, and $X \subseteq Q_v$ is such that $|X| = t^+(A)+1,\ X_2 \subset X$, then $|X_1 \cup \{X \setminus X_2\}| \leqslant t^+(A)$ and $X_1 \cup \{X \setminus X_2\} \in Q(A, A\mathbf{X})$.) Thus, for any matrix A one can determine by a syndrome $\mathbf{Y}$

whether the number of active items of an unknown X is less than $t^+(A)+1$ or not. Later, we shall give another interpretation of $t^+(A)$ and use it to help us design good A's for two-stage testing in certain scenarios. Two-stage algorithms are used in biological applications such as monoclonal antibody generation and cDNA library screening [3,4].

Now we analyze the information which can be extracted from a matrix A of size $b \times v$ and a column vector $\mathbf{Y}$ of length b about the above-defined set $Q(A,\mathbf{Y})$ of those $X \subseteq N_v$ with syndrome $\mathbf{Y}$. There is such a matrix A and syndrome $\mathbf{Y}$ after each stage of an adaptive algorithm, including the first stage. We call an item $j \in N_v$ *negative* if there exists a pool A_i for which $j \in A_i$ and the corresponding test is negative, i.e., $y_i = 0$. We call an item $j \in N_v$ *positive* if there exists a pool A_i for which $j \in A_i$, the set $A_i \setminus \{j\}$ either is empty or consists entirely of negative items, and the corresponding test is positive, i.e., $y_i = 1$. The remaining items will be referred to as *unresolved*. From this definition it follows that any unresolved item j is such that either (i) j does not belong to any of the pools A_i, $i = 1,\dots,b$, or (ii) any A_i such that $j \in A_i$ contains at least one more item which is not negative. Denote the number of negative, positive and unresolved items, respectively, by $n(A,\mathbf{Y})$, $p(A,\mathbf{Y})$, and $u(A,\mathbf{Y})$ and note that

$$n(A,\mathbf{Y}) + p(A,\mathbf{Y}) + u(A,\mathbf{Y}) = v. \tag{4}$$

Example 1. In the following case:

x_1	x_2	x_3	x_4	x_5	x_6	$\mathbf{Y}$
0	1	1	0	0	1	0
1	0	1	1	0	0	1
1	1	1	0	0	0	1

the second, third, and sixth items are negative, the first one is positive, and the fourth and fifth items are unresolved.

It is clear that any negative item must be inactive and any positive item must be active in any set $X \in Q(A,\mathbf{Y})$. Now we verify that unresolved items also justify their name, because $Q(A,\mathbf{Y})$ contains for each unresolved item, u, both a nonempty subset in every member of which u appears and a nonempty subset in every member of which u does not appear.

Lemma 1. *For some A and $\mathbf{Y}$ such that $Q(A,\mathbf{Y}) \neq \emptyset$, let U be the set of unresolved items, W be the set of positive items, and $X_0 = U \cup W$. Then $Q(A,\mathbf{Y})$ contains X_0 and, if U is not empty, $Q(A,\mathbf{Y})$ also contains $X_0 \setminus \{j\}$ for each $j \in U$.*

Proof. By the definitions above, all pools containing unresolved and/or positive items must be positive and all other pools must be negative. This implies that $A\mathbf{X}_0 = \mathbf{Y}$ and hence $X_0 \in Q(A,\mathbf{Y})$. (In particular, if $\mathbf{Y}$ is the all zero syndrome, then $Q(A,\mathbf{Y})$ contains the empty set X_0 which corresponds to the all zero $\mathbf{X}$.) Moreover, all tests do not change their values if we remove any single unresolved item from the set X_0, since a pool containing an unresolved item must contain at least one more nonnegative item. □

Note that, if the unknown X is required to belong to a certain subset $Q \subset Q(v)$, X_0 might not belong to Q in which case we can use this circumstance to determine some unresolved items. In particular, we do this later in the case $Q = Q_t^-(v)$.

Lemma 2. *Let A' be obtained from the matrix A by adjoining one test and $\mathbf{Y}'$ be obtained from the syndrome $\mathbf{Y}$ by adjoining one unit coordinate. If $u(A,\mathbf{Y}) \geqslant 1$, then*

$$u(A,\mathbf{Y}) - 1 \leqslant u(A',\mathbf{Y}') \leqslant u(A,\mathbf{Y})$$

with equality in the left-hand side if and only if the additional pool comprises exactly one of the $u(A,\mathbf{Y})$ unresolved items and a set (possibly empty) of negative items.

Proof. Since the additional test is positive, the number of negative items does not change and positive items remain positive. An unresolved item becomes positive if and only if the additional pool contains this item together with a subset of negative items. □

Note that the set $X = N_v$ gives the all-ones syndrome for any A, since we assumed that no pool is empty. Hence, negative items are absent if the syndrome is all ones. Moreover, we can consider that the initial conditions before the first test are that all items are unresolved. The first test can decrease the number of unresolved items by one only if it is individual. This yields the following perhaps counter-intuitive result about the set $Q_{v-1}^+(v)$ of cardinality $v+1$.

Corollary 1. *There does not exist an adaptive testing algorithm which reconstructs each $X \in Q_{v-1}^+(v)$ based on $v-1$ or fewer tests. Individual testing is the unique algorithm which handles $Q_{v-1}^+(v)$ with v or fewer tests.*

Thus, for any adaptive testing algorithm there exist subsets of $Q(v)$ whose solution requires at least v tests. However, these subsets are not “typical” in a probabilistic sense, and many interesting subsets can be identified on the basis of a much smaller number of tests.

We now consider in more detail *two-stage testing* which in the first stage applies a matrix A of size $b \times v$ to an unknown $X \in Q(v)$ and in the second stage tests individually each of the $u(A,A\mathbf{X})$ unresolved items [4]. Note that if $X = N_v$ and A does not contain individual tests (i.e., rows with one unit), then $u(A,A\mathbf{X}) = v$ and all b tests of the first stage are ineffective because they do not decrease the number of unresolved items. However, if a probability distribution $P(X)$ is given on $X \in Q(v)$, the average number of tests, namely

$$E(A,P) = b + \tilde{u}(A,P), \tag{5}$$

where

$$\tilde{u}(A,P) = \sum_{X \in Q(v)} u(A,A\mathbf{X})P(X),$$

can be much smaller than v. This gives rise to the problem of finding a matrix A which minimizes (5) for a given probability distribution $P(X)$ on the sets $X \in Q(v)$.

Consider a Bernoulli p-scheme in which each item is active with a probability p, $0 < p < 1$, and is inactive with the probability $q = 1 - p$ independent of others. In general, we believe that p might be a nonincreasing function of v and write $p = p(v)$, e.g., p is a constant, or $p = v^{-1/2}$, or $p = v^{-1}$. For the Bernoulli p-scheme, $P(X) = p^i q^{v-i}$ if $|X| = i$, and we use notations $E(A, p)$ and $\tilde{u}(A, p)$ for the corresponding values in (5). For given functions $p = p(v)$ and $b = b(v)$, we call a sequence of matrices $A = A(v)$ of size $b(v) \times v$ *asymptotically good* if $E(A, p)/v \to 0$ as $v \to \infty$.

Together with $\tilde{u}(A, p)$ consider also the following mean values:

$$\tilde{n}(A, p) = \sum_{i=0}^{v} \sum_{X \in Q(v),\ |X| = i} n(A, A\mathbf{X}) p^i q^{v-i}$$

and

$$\tilde{p}(A, p) = \sum_{i=0}^{v} \sum_{X \in Q(v),\ |X| = i} p(A, A\mathbf{X}) p^i q^{v-i}.$$

Lemma 3. *For any $b \times v$ matrix A,*

$$\tilde{n}(A, p) \leqslant \frac{b}{ep}, \quad \tilde{p}(A, p) \leqslant \frac{b}{eq}, \tag{6}$$

$$\tilde{u}(A, p) \geqslant v - \frac{b}{epq}. \tag{7}$$

Proof. Denote by k_i the number of ones in the row $\mathbf{A}_i$ of the matrix A. For any item $j \in N_v$, its probability of being negative (positive) does not exceed $\sum_{i \in B_j} q^{k_i}$ (respectively, $p \sum_{i \in B_j} q^{k_i - 1}$). Therefore,

$$\tilde{n}(A, p) \leqslant \sum_{j=1}^{v} \sum_{i \in B_j} q^{k_i} = \sum_{i=1}^{b} k_i q^{k_i},$$

$$\tilde{p}(A, p) \leqslant \frac{p}{q} \sum_{i=1}^{b} k_i q^{k_i}$$

and hence by (4)

$$\tilde{u}(A, p) \geqslant v - \frac{1}{q} \sum_{i=1}^{b} k_i q^{k_i}. \tag{8}$$

This completes the proof, since the function xq^x has maximum at $x = -1/(\ln q)$ where it equals $1/(-e \ln q)$ and does not exceed $1/(ep)$. □

Corollary 2. *There does not exist an asymptotically good sequence of matrices* $A = A(v)$ *if* $p = p(v)$ *is a constant* $(0 < p < 1)$.

Proof. The corollary follows from (5) and (7); note in this regard from (5) that b/v must vanish as $v \to \infty$ in order for asymptotic goodness to prevail. □

Corollary 2 is not surprising to information theorists. They know that the entropy $H(X)$ of the unknown v-dimensional binary random vector X equals $vh(p)$ bits for the Bernoulli p-scheme, where $h(p)$ is Shannon's entropy function, $h(x) = -x\log_2 x - (1-x)\log_2(1-x)$. Since each test has a binary outcome, learning the result of a test can reduce the uncertainty by at most one bit. It follows that any testing procedure, even one not limited to two stages of testing, must conduct at least $vh(p)$ tests in order to fully resolve X. Since $h(p) > 0$ for $0 < p < 1$, no asymptotically good testing procedures exists for any constant $p \in (0,1)$, not even if the number of stages is allowed to tend to infinity as $v \to \infty$.

Corollary 3. *If* $A = A(v)$ *is an asymptotically good sequence of matrices, then* $k(A) = \max_{1 \leqslant i \leqslant b} k_i \to \infty$ *as* $v \to \infty$.

Proof. Since xq^x increases with x if $x < 1/(-e\ln q)$ using (8) we have

$$\tilde{u}(A,p) \geqslant v - bk(A)q^{k(A)-1} \text{ if } k(A) \leqslant -\frac{1}{\ln(1-p)}.$$

This completes the proof because $p \to 0$ by Corollary 2. □

Note that Corollary 3 prevents one from using "low-density" matrices A to construct an efficient two-stage testing procedure for large v and small $p(v)$.

Two pools are referred to as *noncomparable* if neither of them is a subset of the other.

Lemma 4. *Among matrices* A *which minimize* (5), *there exists a matrix for which all pools are pairwise noncomparable and contain at least two items if this minimum is less than* v.

Proof. We can assume that A does not contain identical or zero rows, since, otherwise, one can remove a row without changing the number of unresolved items. Suppose $A_i \subset A_k$ in violation of noncomparability. Note that $|A_k| \geqslant 2$ because A_i is not empty. Let $j \in A_i$ be such that $a_{i,j} = a_{k,j} = 1$. Consider the matrix A' obtained from A by setting $a_{k,j} = 0$ and leaving all other entries unchanged. We will show that for any $\mathbf{X}$,

$$u(A', A'\mathbf{X}) \leqslant u(A, A\mathbf{X}), \tag{9}$$

which in view of (5) implies that $E(A',p) \leqslant E(A,p)$. Indeed, any pool A'_h, $h = 1,\ldots,b$, of the matrix A' will be negative for the syndrome $\mathbf{Y}' = A'\mathbf{X}$ if A_h is negative for $\mathbf{Y} = A\mathbf{X}$. It follows that any item $l \in N_v \setminus \{j\}$ will be negative for $\mathbf{Y}' = A'\mathbf{X}$ if it is negative for $\mathbf{Y}=A\mathbf{X}$. This is also true for the item j, because if j is negative for $\mathbf{Y}=A\mathbf{X}$, then it belongs to at least one negative pool different from A_k (here we use $A_i \subset A_k$, $a_{i,j}=a_{k,j}=1$, and hence A_i is negative pool if A_k is so). Since all negative items for $\mathbf{Y}=A\mathbf{X}$ remain negative for $\mathbf{Y}'=A'\mathbf{X}$, all positive items remain positive with the possible exception of j when the pool A_k is positive and all its items, different from j, are negative. However, $A_i \subset A_k$ implies that in this case the pool A_i' will be positive and all its items different from j (if they exist) will be negative. Therefore, the positive item j remains positive as well. Thus, no item's status as negative or positive can change if one switches from A to A' and the number of unresolved items can be only decreased; this proves (9). Finally, if $|A_i|=1$, $a_{i,j}=1$ and all pools are pairwise noncomparable, then $a_{l,j}=0$ for all $l \neq i$. Therefore, if we remove ith row of A, then the number of unresolved items increases by 1 and the sum (5) is not changed. However, all b rows cannot have this property because in this case A has $v-b$ zero columns and hence $E(A,p)=v$. □

The following example shows that removing from one of a pair of noncomparable pools an item that is common to both these pools can increase the number of unresolved items for a certain X; this does not contradict Lemma 4.

Example 2.

X	1	0	0	1	0	**Y**
	1	1	1	0	0	1
	1	1	0	1	0	1
	0	1	0	0	1	0
	0	0	1	0	1	0
	p	n	n	u	n	

X	1	0	0	1	0	**Y**
	0	1	1	0	0	0
	1	1	0	1	0	1
	0	1	0	0	1	0
	0	0	1	0	1	0
	u	n	n	u	n	

3. Using cover-free codes in two-stage testing

For a matrix A of size $b \times v$ with subsets $B_j \subseteq N_b$, $j=1,\dots,v$, of the numbers of the unit coordinates of its columns and a set $X \in Q(v)$, we define the *closure* $\bar{X}$ of X as follows:

$$\bar{X}=\{j \in N_v : B_j \subseteq Y\} \quad \text{where } \mathbf{Y}=A\mathbf{X}. \tag{10}$$

From this definition it follows that $\bar{X}$ consists of all unresolved and positive items and does not contain negative items for the syndrome $\mathbf{Y}$. If $X=\{j_1,\dots,j_t\}$, $t \geqslant 1$, and $j \in \bar{X} \setminus X$, then j is an unresolved item and

$$B_j \subseteq \bigcup_{h=1}^{t} B_{j_h},$$

i.e., B_j is *covered* by $\bigcup_{h=1}^{t} B_{j_h}$. Herewith an item $j \in X$ is unresolved or positive depending on whether $\overline{X \setminus \{j\}}=\bar{X}$ or not. If $\bar{X}=X$ for any $X \in Q_t^-(v)$, then the set of columns of A is called a *cover-free* (v,b,t)-code (or t-cover-free code). For any A

denote by $t(A)$ the maximum number t such that $\overline{X}=X$ for any $X\subseteq N_v$ if $|X|\leqslant t$. Disjunctive and cover-free (v,b,t)-codes were introduced in [12]. In [12] it was also shown that

$$t(A)\leqslant t^{-}(A)\leqslant t(A)+1. \tag{11}$$

This follows from the facts that if $A\mathbf{X}_1=A\mathbf{X}_2$ where $X_1\neq X_2$, then $\overline{X_1}\neq X_1$ or $\overline{X_2}\neq X_2$, and if $j\in\overline{X}\setminus X$, then $A\mathbf{X}=A\mathbf{X}_1$ where $X_1=X\cup\{j\}$.

In this section we use the fact that cover-free codes have an important property which is useful for their application as the first stage of two-stage testing (this property was also considered in [2] in a more general content). t-cover-free codes not only allow us to recover the unknown vector from its syndrome if the number of its active items does not exceed t, but also allow us to determine from this syndrome whether the number of active items exceeds t or not. This makes it possible to construct an efficient two-stage test for a given probability distribution P the first stage of which is a t-cover-free code with t slightly larger than the expected number of active items.

Lemma 5. *For any $b\times v$ matrix A*

$$t^{+}(A)=t(A), \tag{12}$$

and for any probability distribution P on $Q(v)$,

$$E(A,P)\leqslant b+\sum_{X\in Q(v),|X|>t(A)}u(A,A\mathbf{X})P(X)$$

$$\leqslant b+v\sum_{X\in Q(v),|X|>t(A)}P(X).$$

Proof. If there exist $X_1\in Q_t^{-}(v)$ and $X_2\in Q_{t+1}^{+}(v)$ such that $A\mathbf{X}_1=A\mathbf{X}_2$, then $\overline{X_1}=\overline{X_2}$ and $\overline{X_2}$ contains at least one element which does not belong to X_1 and hence $\overline{X_1}\neq X_1$. This implies $t^{+}(A)\geqslant t(A)$. On the other hand, let $j\in\bar{X}\setminus X$ for some $X\in Q_t^{-}(v)$ and let X_1 be a set of size t such that $X\subset X_1$ and $j\notin X_1$. Then for $X_1\in Q_t^{-}(v)$ and $X_2=X_1\cup\{j\}\in Q_{t+1}^{+}(v)$, we have $A\mathbf{X}_1=A\mathbf{X}_2$, which shows that $t(A)\geqslant t^{+}(A)$, so property (12) is established. When using a two-stage testing algorithm, this property of the matrix A allows us to determine from a syndrome of an unknown X whether $|X|\leqslant t(A)$ or not. In the first case we use the fact that $t^{-}(A)\geqslant t(A)$ (or a cover-free (v,b,t)-code is a disjunctive (v,b,t)-code) to reconstruct this X. In the second case we apply individual tests to all unresolved items of X. This completes the proof. □

Denote by $b(v,t)$ the minimum number of rows in a binary matrix A with v columns which form a t-cover-free code. Upper bounds on $b(v,t)$ have been obtained using both random test selection and known error-correcting codes (see [8,10–12]). We use below the best known asymptotic upper bound obtained in [9] (see also [14]):

$$b(v,t)\lesssim(t\log_2 e)^2\ln v\quad\text{as }v\to\infty,\ t\to\infty. \tag{13}$$

Using Lemma 5 for a Bernoulli p-scheme we estimate $E(A,p)$ for a t-cover-free code A with $t=t(A)=vp_0$ where $p_0>p$. Let

$$f(x)=x\ln\frac{x}{e}+1.$$

For any y, $0<y<\infty$, the equation $f(x)=y$ has a unique solution $\mu=\mu(y)>1$. Note that

$$\mu(y)\sim 1 \quad \text{if } y\to 0 \tag{14}$$

and

$$\mu(y)\sim \frac{y}{\ln y} \quad \text{if } y\to\infty. \tag{15}$$

Theorem 1. *There exist t-cover-free matrices $A=A(v)$ of size $b(v)\times v$ with*

$$t=\left\lfloor vp\mu\left(\frac{\ln v}{vp}\right)\right\rfloor \tag{16}$$

such that $\tilde{u}(A,p)\leqslant 1$ and

$$b(v)\lesssim \left(vp(\log_2 e)\mu\left(\frac{\ln v}{vp}\right)\right)^2\ln v \quad \textit{as } v\to\infty. \tag{17}$$

Proof. By Lemma 5 and (13), it is sufficient to show that the integer t defined by (16) tends to ∞ as $v\to\infty$ and

$$v\sum_{i\geqslant t+1}\binom{v}{i}p^i(1-p)^{v-i}\leqslant 1. \tag{18}$$

Put $z=p\mu((\ln v)/(vp))$ and note that $z>p$ and $t+1>vz$. Using the Chernoff bound and a standard inequality we get

$$\begin{aligned} &v\sum_{i\geqslant t+1}\binom{n}{i}p^i(1-p)^{v-i}\\ &\quad\leqslant v\exp\left\{v\left(-z\ln\frac{z}{p}+(1-z)\ln\left(1+\frac{z-p}{1-z}\right)\right)\right\}\\ &\quad\leqslant v\exp\left\{v\left(-z\ln\frac{z}{ep}-p\right)\right\}=\exp\left\{\ln v-vpf\left(\frac{z}{p}\right)\right\}=1. \end{aligned}$$

Since $\mu(y)>1$, we have $t\to\infty$ if $vp\to\infty$. If vp is restricted and hence $(\ln v)/(vp)\to\infty$ as $v\to\infty$, then $t\sim(\ln v)/(\ln\ln v)$, by (15). □

We give the following special cases of (17) which follow from (14) and (15):

$$b(v) \lesssim (vp\log_2 e)^2 \ln v \quad \text{if} \quad \frac{\ln v}{vp} \to 0, \tag{19}$$

$$b(v) \lesssim (\log_2 e)^2 \, \frac{(\ln v)^3}{(\ln((\ln v)/vp))^2} \quad \text{if} \quad \frac{\ln v}{vp} \to \infty.$$

In the case $p = (\ln v)/v$ we have $\mu((\ln v)/(vp)) = \mu(1) = e$ and (17) gives

$$m(v) \lesssim (e\log_2 e)^2 (\ln v)^3.$$

In particular, from (19) it follows that there exist asymptotically good sequences based on cover-free codes when $p = p(v) = \mathrm{o}(1/\sqrt{v\ln v})$.

4. Combinatorial designs as the first stage of testing

In combinatorial theory, k-subsets of the set $N_v = \{1, 2, \ldots, v\}$ are commonly referred to as *blocks*. A set S of blocks is called a $2-(v,k,1)$ *design*, or *Steiner 2-design*, if any two different elements of N_v belong to one and only one block. From this definition it follows that if b is the number of blocks in the Steiner 2-design, then

$$bk(k-1) = v(v-1) \tag{20}$$

and each element of N_v belongs to the same number

$$r = \frac{v-1}{k-1} \tag{21}$$

of blocks. These equalities give some necessary arithmetic conditions for existence of $2-(v,k,1)$ designs, which are sufficient for fixed k and sufficiently large v by the Wilson theorem (see, for example, [7]). The best-known infinite family of Steiner 2-designs consists of $2-(v,3,1)$ designs (the Steiner triples) which exist if and only if v has the form $6l+1$ or $6l+3$, $l = 1, 2, \ldots$. The Fisher inequality $b \geqslant v$ holds for Steiner 2-designs, and this fact disables using their blocks as pools for testing. In this connection we weaken the conditions imposed on Steiner 2-designs so that the opposite inequality $b \leqslant v$ may hold.

Consider a matrix $A = (a_{i,j})$ of size $b \times v$ with entries 0 and 1. As before we denote by $\mathbf{A}_i$ the ith row of A and by A_i the pool which is the subset of N_v consisting of the coordinate positions of the nonzero entries of $\mathbf{A}_i$. Analogously, we denote by $\mathbf{B}_j$ the jth column of A and by B_j the subset of N_b which consists of the coordinate positions of the nonzero entries of $\mathbf{B}_j$ and is called a *block*. (We shall use below blocks of Steiner 2-designs as blocks of matrices A.) Such a matrix A will be referred to as a (v,k,b,r)*-design*, if each pool is a k-set, each block is an r-set, and any two different items of N_v belong to at most one pool. In particular, the Steiner $2-(v,k,1)$ designs form a subclass of (v,k,b,r)-designs.

Another possible interpretation of a (v,k,b,r)-design A is based on a bipartite graph whose parts consist of b and v vertices such that a vertex i of the first part is adjacent to a vertex j of the second part if and only if $a_{i,j}=1$. The conditions in the definition of a (v,k,b,r)-design mean that the degree of any vertex of the first part equals k, the degree of any vertex of the second part equals r, and the bipartite graph does not have any cycles of length 4 or less.

From the definition of a (v,k,b,r)-design it follows that

$$vr = bk \tag{22}$$

and that the transpose A^{T} of A forms a (b,r,v,k)-design. For any a (v,k,b,r)-design A, we have the two inequalities

$$bk(k-1) \leqslant v(v-1) \tag{23}$$

and

$$r(k-1) \leqslant v-1. \tag{24}$$

(23) is proved, similarly to (20), using a count of the total number of pools containing all pairs of elements, and (24) follows from (23) and (22). Equality prevails in each of these inequalities if and only if A is a $2-(v,k,1)$ design. Since for any (v,k,b,r)-design A the matrix A^{T} forms a (b,r,v,k)-design, we also have two other inequalities,

$$vr(r-1) \leqslant b(b-1) \tag{25}$$

and

$$k(r-1) \leqslant b-1 \tag{26}$$

in which equality holds if and only if A^{T} is a $2-(b,r,1)$ design.

Any (v,k,b,r)-design A can be considered as a constant-weight code of size b in the Johnson space J_v^k, which consists of binary vectors of Hamming weight k and length v; the preferred measure of distance in J_v^k, called the Johnson distance, equals half the Hamming distance between vectors. Since this code has the same number r of ones in any column, it is a 1-design in the Johnson space (in the terminology of Delsarte [6]). Moreover, from the definition of a (v,k,b,r)-design A it follows that at most two different Johnson distances between different rows of A are possible, namely $k-1$ and k. By Delsarte's theorem [6] for association schemes, since this code is a 1-design with at most two distances, it must be *distance-invariant*. That is, there exist two numbers b_{k-1} and b_k such that for any row $\mathbf{A}_i$ of A, $i=1,\ldots,b$, the number of rows at distance $k-1$ and k are equal to b_{k-1} and b_k, respectively. Each column has r ones and hence the total sum of the Johnson distances between ordered pairs of rows equals $vr(b-r)=bk(b-r)$. Since this sum is also equal to $b((k-1)b_{k-1}+kb_k)$, and $b_{k-1}+b_k=b-1$, we have

$$b_{k-1}=k(r-1) \quad \text{and} \quad b_k=b-1-k(r-1). \tag{27}$$

A similar statement is true for the code formed by v columns of the matrix A which have r ones and $b-r$ zeros.

There are five known infinite families of $2-(b,r,1)$ designs (see [7]). We list the parameters of the corresponding families of (v,k,b,r)-designs which are obtained by transposition of these Steiner systems

$$\left(\frac{b(b-1)}{6}, \frac{b-1}{2}, b, 3\right), \tag{28}$$

where $b = 6l + 1$ or $b = 6l + 3$, $l = 1, 2, \ldots$;

$$\left(m^2\frac{m^3+1}{m+1}, m^2, m^3+1, m+1\right), \tag{29}$$

$$\left(m^{n-1}\frac{m^n-1}{m-1}, \frac{m^n-1}{m-1}, m^n, m\right), \tag{30}$$

$$\left(\frac{(m^n-1)(m^{n+1}-1)}{(m-1)(m^2-1)}, \frac{m^n-1}{m-1}, \frac{m^{n+1}-1}{m-1}, m+1\right), \tag{31}$$

$$((2^s+1)(2^s-2^{s-l}+1), 2^s+1, 2^{s+l}-2^s+2^l, 2^l), \tag{32}$$

in (29)–(31) m is a prime power and $n = 2, 3, \ldots$ (n-dimensional affine and projective geometries over GF(m) are used in these constructions), in (32) s and l are any integers such that $s > l \geqslant 2$ (Denniston designs).

Now we investigate the property of (v, k, b, r)-designs A ($k \geqslant 2$, $r \geqslant 2$) for the first stage of testing. Note that (25) shows that the number b of tests cannot be too small, and this estimate is attained for all designs (28)–(31). As was already remarked in [12], $t(A) = r - 1$. This is true, since for $k \geqslant 2$ each block B_j is covered by the union of some r blocks $B_{j(h)}$, $h = 1, \ldots, r$, where $j, j(1), \ldots, j(r)$ all are different, and this number r cannot be decreased because any two different blocks have at most one common element. By (12) we have

$$t^+(A) = t(A) = r - 1. \tag{33}$$

Thus, by using a (v, k, b, r)-design A at the first stage, one can determine from a syndrome of an unknown $X \in Q(v)$ whether the number of its active items exceeds $r - 1$ or not and reconstruct $X \in Q(v)$ in the first case without additional individual tests.

Lemma 6. *For any (v, k, b, r)-design A such that $b < v$,*

$$r - 1 \leqslant t^-(A) \leqslant r \leqslant \sqrt{b}. \tag{34}$$

Proof. Since $t(A) = r - 1$, the inequalities $r - 1 \leqslant t^-(A) \leqslant r$ follow from (11). The Johnson distance between different columns of A is not less than $r - 1$. This implies the inequality $r \leqslant \sqrt{b}$ because, by a recent result [1], the size of a code consisting of vectors with r ones and $b - r$ zeros and having the minimal Johnson distance d does not exceed b if $d > r(b - r)/b$. □

The following statement refines (11) and (34) in a special case.

Lemma 7. *If a (v,k,b,r)-design A is such that A^{T} forms a $2-(b,r,1)$ design, then*

$$t^{-}(A)=t(A)=t^{+}(A)=r-1.$$

Proof. If A^{T} is a $2-(b,r,1)$ design, then for any two elements of N_b there exists one (and only one) block containing these elements. We shall use this property to find $r+1$ different blocks $B_j, B_{j'}, B_{j(1)}, \ldots, B_{j(r-1)}$ such that

$$B_j \cup \cup_{h=1}^{r-1} B_{j(h)} = B_{j'} \cup \cup_{h=1}^{r-1} B_{j(h)}. \tag{35}$$

To do this, take as B_j and $B_{j'}$ any two different blocks having a common element, and partition their remaining elements into $r-1$ pairs so that any pair does not belong to one of these blocks. As $B_{j(1)}, \ldots, B_{j(r-1)}$ we choose blocks containing these $r-1$ pairs. These blocks must differ from B_j and $B_{j'}$ and be distinct for, if they were to coincide, their common value would be a block that has two elements in common with B_j and $B_{j'}$. Since (35) shows that the sets $\{j, j(1), \ldots, j(r-1)\}$ and $\{j', j(1), \ldots, j(r-1)\}$ give the same syndrome, the proof is complete. □

Note that the inequality $r \leqslant \sqrt{b}$ is also attained for (v,k,b,r)-designs (30) when $n=2$.

Theorem 2. *Given a (v,k,b,r)-design A and a Bernoulli model with the parameter p,*

$$v(1-q^{k-1})^r \leqslant \tilde{u}(A,p) \leqslant v(p+q(1-q^{k-1})^r). \tag{36}$$

Proof. Consider an arbitrary item $j \in N_v$. There are r pools containing j. Each of them has size k, and item j is the only point of intersection of any pair of them. Therefore, the probability of the event that each of these r pools contains at least one active item among its $k-1$ remaining elements equals $(1-q^{k-1})^r$. Regardless of whether item j is active or inactive, it will be unresolved because all these pools are positive in this case. This gives the lower bound in (36). To prove the upper bound we use the fact that

$$\tilde{u}(A,p) \leqslant \sum_{j \in N_v} \sum_{B_j \subseteq AX} P(X). \tag{37}$$

Considering two cases when j is active j is inactive but all pools A_i such that $a_{i,j}=1$ are positive, we find that the right-hand side of (37) equals $p+q(1-q^{k-1})^r$. □

Example 3. Consider a (v,k,b,r)-design A given by (28), for instance, the $(100,12,25,3)$-design A obtained from the Steiner triple 2-$(25,3,1)$. Let $\mathbf{Y}=A\mathbf{X}$. We have $|Y| \leqslant 6$ if $|X| \leqslant 2$ and $|Y| \geqslant 7$ if $|X| \geqslant 4$. In the case $|Y|=6$ we have two possibilities, $|X|=2$ or $|X|=3$. However, since $t^{+}(A)=2$ we can distinguish these two cases and also find X if $|X|=2$. For the $(100,12,25,3)$-design A and $p=0.01$, the upper bound in (36) shows that $\tilde{u}(A,p)<1.1135$.

Theorem 2 and Corollary 2 imply the following.

Corollary 4. *A sequence of* (v, k, b, r)*-designs* A *is asymptotically good if and only if* $p = p(v) \to 0$*,* $b/v \to 0$*, and*

$$(1 - (1 - p)^{k-1})^r \to 0 \quad \textit{as } v \to \infty. \tag{38}$$

In general, we cannot state that $\tilde{u}(A, p)$ grows with p for any matrix A. However, $(1-(1-p)^{k-1})^r$ is an increasing function of p; hence, a sequence of (v, k, b, r)-designs that is asymptotically good for $p = p(v) \to 0$ also is asymptotically good for any smaller $p = p(v)$.

Theorem 3. *Sequence* (28) *of* $(v, k, b, 3)$*-designs* A *is asymptotically good if and only if* $p = p(v) = \mathrm{o}(v^{-1/2})$*. Moreover,*

$$E(A, p) \sim \sqrt{6v} \quad \textit{if} \;\; p = p(v) = \mathrm{o}(v^{-2/3}). \tag{39}$$

Proof. For the sequence (28) we have $r = 3$ and $k \sim \sqrt{3v/2}$, $b \sim \sqrt{6v}$ as $v \to \infty$. From (38) it follows that this sequence is asymptotically good if and only if $q^{k-1} \to 1$, which is equivalent to the condition $pk \to 0$. This proves the first part of the statement. The second part is a consequence of (5), (36), and the fact that if $pk \to 0$, then $(1 - q^{k-1})^r \sim (pk)^3$. □

Theorem 4. *If* $p = p(v) \leqslant c(\ln v)/\sqrt{v}$ *where* $c < \frac{1}{4}$*, then the sequence* (29) *of* (v, k, b, r)*-designs* A *(and sequence* (30) *or* (31) *with* $n = 3$*) is asymptotically good and*

$$E(A, p) \sim v^{3/4} \quad \textit{as } v \to \infty. \tag{40}$$

If $p = p(v) \geqslant c\,(\ln v)/\sqrt{v}$ *where* $c > \frac{1}{4}$*, then none of these sequences is asymptotically good.*

Proof. Note that for each sequence we have $k \sim v^{1/2}$, $b \sim v^{3/4}$, and $r \sim v^{1/4}$ as $v \to \infty$. Therefore, if $p = p(v) = c(\ln v)/\sqrt{v}$, $c > 0$, then

$$(1 - p)^{k-1} = v^{-c(1+\mathrm{o}(1))}$$

and

$$(1 - (1 - p)^{k-1})^r \sim \exp\{-v^{1/4-c+\mathrm{o}(1)}\}.$$

It follows that we have (38) and (40) if $c < \frac{1}{4}$, whereas we have $(1 - q^{k-1})^r \to 1$ if $c > \frac{1}{4}$, which completes the proof. □

In [5] the authors recently found the order of magnitude of

$$E(v, p) = \min E(A, p)$$

where the minimum is taken over *all* matrices A with v columns as $v \to \infty$ and $p \to 0$. This result uses lower and upper bounds on $\tilde{u}(A, p)$ obtained via random selection and the linear programming approach suggested by Knill in [13].

References

[1] E. Agrell, A. Vardy, K. Zeger, Upper bounds for constant-weight codes, IEEE Trans. IT 46 (2000) 2373–2395.

[2] D.J. Balding, D.C. Torney, Optimal pooling designs with error detection, J. Combin. Theory, A 74 (1996) 131–140.

[3] E. Barillot, B. Lacroix, D. Cohen, Theoretical analysis of library screening using an N-dimensional pooling strategy, Nucleic Acids Res. 19 (1991) 6241–6247.

[4] T. Berger, V.I. Levenshtein, Asymptotic efficiency of two-stage disjunctive testing, IEEE Trans. IT 48 (7) (2002) 1741–1749.

[5] T. Berger, J.W. Mandell, P. Subrahmanya, Maximally efficient two-stage group testing, Biometrics 56 (2000) 107–114.

[6] P. Delsarte, An algebraic approach to the association schemes of coding theory, Philips Res. Rep. Suppl. 10 (1973).

[7] J.H. Dinitz, D.R. Stinson (Eds.), Contemporary Design Theory, Wiley-Interscience, New York, 1992.

[8] A.G. Djachkov, V.V. Rykov, A survey of superimposed distance codes, Problems Control Informat. Theory 12 (1983) 11–13.

[9] A.G. Djachkov, V.V. Rykov, On constant-weight disjunctive codes, in: Proceedings of IX Symposium on Redundancy Problem in Information Systems, Part I, Leningrad, 1986, pp. 116–119 (in Russian).

[10] D.-Z. Du, F.K. Hwang, Combinatorial Group Testing and its Applications, World Scientific, Singapore, 1993.

[11] F.K. Hwang, V.T. Soś, Non-adaptive hypergeometric group testing, Stud. Sci. Math. Hung. 22 (1987) 257–263.

[12] W.H. Kautz, R.R. Singleton, Nonrandom binary superimposed codes, IEEE Trans. IT 10 (4) (1964) 363–377.

[13] E. Knill, Lower bounds for identifying subset members with subset queries. in: Proceedings of the Sixth Annual ACM–SIAM Symposium on Discrete Algorithms, San Francisco, CA, January 22–24, 1995, 369–377, (Chapter 40).

[14] Q.A. Nguyen, T. Zeisel, Bounds on constant weight binary superimposed codes, Probl. Control Inform. Theory 17 (1988) 223–230.