[10] A. M. Chan and F. R. Kschischang, "A simple taboo-based soft-decision decoding algorithm for expander codes," *IEEE Commun. Lett.*, vol. 2, no. 7, pp. 183–185, Jul. 1998.

[11] J. Zhang and M. P. C. Fossorier, "A modified weighted bit-flipping decoding of low-density parity check codes," *IEEE Commun. Lett.*, vol. 8, no. 3, pp. 165–167, Mar. 2004.

[12] A. Nouh and A. Banihashemi, "Bootstrap decoding of low-density parity check codes," *IEEE Commun. Lett.*, vol. 6, no. 9, pp. 391–393, Sep. 2002.

[13] Z. Liu and D. Pados, "Low complexity decoding of finite geometry LDPC codes," *IEEE Trans. Commun.*, to be published.

[14] Y. Mao and A. Banihashemi, "Decoding low density parity check codes with probabilistic scheduling," *IEEE Commun. Lett.*, vol. 5, no. 10, pp. 414–416, Oct. 2001.

[15] M. Mihaljevic and J. Golic, "A method for convergence analysis of iterative probabilistic decoding," *IEEE Trans. Inf. Theory*, vol. 46, no. 6, pp. 2206–2211, Sep. 2000.

[16] N. Miladinovic, "Iterative decoding of LDPC and GLDPC codes on BSC," Ph.D. dissertation, Dept. Elec. Eng., Univ. Hawaii at Manoa, , 2005, to be published.

[17] X. Y. Hu, E. Eleftheriou, and D. M. Arnold, "Regular and irregular progressive edge-growth tanner graphs," in *Proc. 2001 Global. Telecommun. Conf.*, vol. 2, San Antonio, TX, Nov. 2001, pp. 995–1001.

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

Single-User Tracing and Disjointly Superimposed Codes

Miklós Csűrös and Miklós Ruszinkó

***Abstract*—The zero-error capacity region of r-out-of-T user multiple-access OR channel is investigated. A family $\mathcal{F}$ of subsets of $[n] = \{1, \dots, n\}$ is an r-single-user-tracing superimposed code (r-SUT) if there exists such a single-user-tracing function $\phi\colon 2^{[n]} \mapsto \mathcal{F}$ that for all $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \leq |\mathcal{F}'| \leq r$, $\phi(\cup_{A \in \mathcal{F}'} A) \in \mathcal{F}'$. In this correspondence, we introduce the concept of these codes and give bounds on their rate. We also consider disjointly r-superimposed codes.**

***Index Terms*—Codes, group testing, physical mapping, superimposed codes.**

I. Introduction

Suppose that T users share a common channel. A binary vector of length n is associated to each user. The ith user transmits its vector $\boldsymbol{x}_i = (x_i^1, x_i^2, \dots, x_i^n)$ $(i = 1, 2, \dots, T)$ if it is active, otherwise it does not. It is assumed that the transmission is bit and block synchronized. The destination of the messages is a single receiver that observes the bitwise OR vector of the vectors

$$\boldsymbol{y} = \bigvee_{\forall i \text{ active}} \boldsymbol{x}_i$$

associated to the active users. Moreover, suppose that at most r users are active simultaneously. In the classical framework of superimposed coding, the receiver has to be able to identify the set of all active users from the output vector $\boldsymbol{y}$ of the channel. That is, the code must satisfy the property that for all choices of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$ and $\boldsymbol{z}_1, \dots, \boldsymbol{z}_\ell$ of codewords with $1 \leq k, \ell \leq r$ and $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_k\} \neq \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_\ell\}$, we have

$$\bigvee_{i=1}^{k} \boldsymbol{x}_i \neq \bigvee_{j=1}^{\ell} \boldsymbol{z}_j.$$

Although the rate of these codes have been studied extensively in, e.g., [1]–[6], it remains to be determined: the gap between the known upper and lower bounds is still substantially large.

Here we investigate the case when the receiver has to be able to identify *just one* user out of at most r active ones. Clearly, if a code is superimposed in the classical sense then it satisfies this requirement: being able to identify all active users, the receiver can always name just one. A practical motivation for studying r-single-user-tracing (r-SUT) families rises from applications of combinatorial designs in genomics, reviewed in Section II. Section III discusses our results on the rate of SUT superimposed codes. Section IV introduces the class of disjointly superimposed codes, and analyzes their extremal properties. Section V concludes the correspondence with some open questions.

II. Superimposed Codes for the Physical Mapping of Genomic Clones

A recently emerging application of superimposed codes, and group testing methods in general, is for the analysis of genomic data. Examples include the quality-control of DNA chips [7], and diverse applications related to genome sequencing: closing the remaining gaps at the end of a sequencing project [8], and clone library screening [9], which we consider here in more detail. The sequencing of large genomes (such as human) rely on *genomic clones*. We describe here briefly the relevant procedures, somewhat simplifying the problem. A recent overview of large genome sequencing techniques is given by Green [10]. The genome of an organism can be described by a sequence over a four-letter alphabet, corresponding to the four nucleotides used in DNA. Mammalian genome sizes are in the order of billions. For our purposes, a genomic clone is a random contiguous fragment of the genome. (Fragments are inserted into a host cell, which multiplies and thus creates many identical copies of the original cell containing the same piece of inserted foreign DNA fragment, hence the term "clone.") Typical clone fragment sizes are 100–200 thousand nucleotides. A *clone library* is a collection of genomic clones, produced using a large number of random fragments from many genome copies. The fragments correspond essentially to a uniform sampling of the whole genome. The information on which part of the genome the fragments originate from is lost in the course of random sampling, and needs to be determined using additional techniques. In a preliminary step to complete genome sequencing, called *physical mapping*, this information is established, by exploring overlaps between clone fragments. Using the physical map, a smaller set of minimally overlapping clones is selected in order to sequence the clones one-by-one. For instance, while sequencing the human genome, more than 300 thousand genomic clones were analyzed and about 30 thousand were selected for complete sequencing [11].

Manuscript received July 29, 2004; revised October 25, 2004. This work was supported by NSERC under Grant 250391-02, and by OTKA under Grants T038198 and T046234. This work relates to Department of the Navy Grant N00014-04-1-4034 issued by the Office of Naval Research International Field Office and to the Air Force Office of Scientific Research, NSAF, Grant FA 8655-05-1-3017. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. AMS classification numbers: 94B65, 94B25, 94A40. The material in this correspondence was presented in part at the IEEE Symposium on Information Theory, Chicago, IL, June/July 2004.

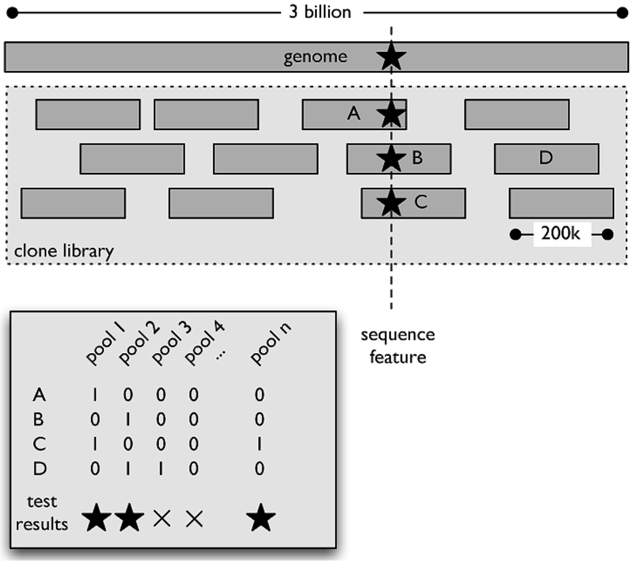
M. Csűrös is with the Department of Computer Science and Operations Research, Université de Montréal, Montréal, QC H3C 3J7, Canada (e-mail: csuros@iro.umontreal.ca).

M. Ruszinkó is with the Computer and Automation Research Institute, the Hungarian Academy of Sciences, Budapest, P. O. Box 63, H-1518, Hungary (e-mail: ruszinko@ lutra.sztaki.hu).

Communicated by A. E. Ashikhmin, Associate Editor for Coding Theory.

Digital Object Identifier 10.1109/TIT.2005.844097

0018-9448/$20.00 © 2005 IEEE



	pool 1	pool 2	pool 3	pool 4 ...	pool n
A	1	0	0	0	0
B	0	1	0	0	0
C	1	0	0	0	1
D	0	1	1	0	0
test results	★	★	×	×	★

Fig. 1. Clone library screening. A clone library is a collection of random fragments from a genome. Clones in a library are tested for the presence of a sequence feature, such as homology to a given region in a related genome. The tests are carried out by pooling the clones: if the clone subset comprising the pool contains the feature, the test is positive, otherwise it is not.

The main issue in constructing a physical map is the discovery of overlaps. The key technique is to test sequence features, which are necessarily shared by overlapping clones. Often, a group-testing approach is employed by *pooling* the DNA from different clones: a pool is defined by a subset of clones that are screened together in a single experimental step. Fig. 1 illustrates the concept of clone pooling. In the terminology of superimposed codes, clones correspond to users and pools correspond to the coordinates of the user vectors: the ith clone is included in pool j if the jth bit of user vector $\boldsymbol{x}_i$ equals one. Active users correspond to clones containing a particular feature. When testing a feature, pools are tested individually, and exactly those pools that contain a clone with the given feature test positive. The set (or at least one) of the clones containing the feature has to be determined from the set of positive pools, in the same way as the set of active users needs to be determined from the bitwise OR of their vectors.

Historically, the most widely used features are short (up to the order of hundreds of letters) contiguous sequences that occur once in the genome, called *sequence tagged sites* (STS). All DNA in a pool can be tested for the presence of a given STS, by hybridization for example. Pooling designs for the purposes of STS screening have been studied extensively [9], [12], [13], and this particular application inspired many recent theoretical results on superimposed codes and nonadaptive group testing procedures [14]–[17].

A more recent application uses shotgun sequences [18], [19] for testing sequence features in pooled clones. Pooled genomic indexing (PGI) [19] maps genomic clones to a reference genome sequence. Thus, the type of sequence feature that is tested by PGI is similarity to a region in the reference genome. In contrast to STS screening, the features are not defined before the experiment but are found in the analysis of the outcome. In a current application, (unsequenced) rhesus macaque clones are being mapped to the human genome. The raw outcome of the experiment is a list of mappings between sets of pools and regions in the reference sequence. Each mapping is indicative of the fact that some clones are similar to the same region in the reference sequence. The set of pools containing those clones is observed by the experimenter, along with the reference region.

The results of STS screening or a PGI experiment can be used to select clones for complete sequencing. If the purpose of the experiment is to identify clones that are particularly interesting and to sequence them completely, a single-user tracing code is more adequate for the pooling design than a "fully" superimposed code. In PGI, for instance, a number of overlapping macaque clones may include the same region that is homologous to a particular human gene: the experimenter will want to identify at least one of those clones for complete sequencing, but there is no need to identify all of them as they convey the same information about the genome.

The bound r on the number of "active users," i.e., the number of clones exhibiting a given feature is determined by the number of clones T. The size of a clone library is characterized by the *coverage*, which equals $c = TL/G$ where L is the average length of a clone, and G is the total genome length. Various aspects of clone overlaps can be studied by modeling the clone positions as arrival times in a Poisson process. For example, the number of clones that include a given position in the genome is a Poisson random variable with expected value c [20]. Clone library coverage values are typically below 10, and are rarely above 20. If unique sequence features are used, then every feature is shared by, say, at most $r = \lceil 2c \rceil$ clones with high probability.

III. Single-User Tracing Superimposed Codes

As the question is rather of a combinatorial nature, we introduce a set terminology. Accordingly, codewords are characteristic vectors of subsets of a set $[n] = \{1, \ldots, n\}$ where $n > 0$, i.e., the subset A corresponds to the binary vector $\boldsymbol{x} = (x^1, \ldots, x^n)$ with $x^i = 1$ if and only if $i \in A$, and *vice versa*.

Throughout the correspondence, we use the de Finetti notation for indicator functions, i.e., $\{\cdots\}$ denotes an event, or its indicator function, depending on the context. We write $f(m) = o(g(m))$ if the sequence $f(m)/g(m) \to 0$ as $m \to \infty$. When the base of the logarithm matters, we use lg to denote binary logarithm.

Definition 3.1: A family $\mathcal{F} \subseteq 2^{[n]}$ is r-superimposed if

$$\bigcup_{i=1}^{k} A_i \neq \bigcup_{j=1}^{\ell} B_j$$

for any

$$\{A_1, A_2, \ldots, A_k\} \neq \{B_1, B_2, \ldots, B_\ell\}$$

$1 \leq k, \ell \leq r$; $A_1, A_2, \ldots, A_k, B_1, B_2, \ldots, B_\ell \in \mathcal{F}$.

We are interested in r-SUT families, defined as follows.

Definition 3.2: A family $\mathcal{F}$ is r-SUT if for all choices of $\mathcal{F}_1, \ldots, \mathcal{F}_k \subseteq \mathcal{F}$ with $1 \leq |\mathcal{F}_i| \leq r$

$$\bigcup_{A \in \mathcal{F}_1} A = \bigcup_{A \in \mathcal{F}_2} A = \cdots = \bigcup_{A \in \mathcal{F}_k} A$$

implies $\cap_{i=1}^{k} \mathcal{F}_i \neq \emptyset$. Equivalently, there exists such an SUT function $\phi\colon 2^{[n]} \mapsto \mathcal{F}$ that for all $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \leq |\mathcal{F}'| \leq r$, $\phi(\cup_{A \in \mathcal{F}'} A) \in \mathcal{F}'$.

The following (folklore) lemma shows that it is enough to consider $k \leq r + 1$ in Definition 3.2.

Lemma 3.3: Let $k \geq r + 1$. Let $S_1, \ldots, S_k$ be a collection of sets, each containing at most r elements. If for all choices of $1 \leq i_1 < \cdots < i_{r+1} \leq k$, $\cap_{j=1}^{r+1} S_{i_j} \neq \emptyset$, then $\cap_{i=1}^{k} S_i \neq \emptyset$.

Proof: For the sake of contradiction, suppose that $\cap_{i=1}^{k} S_i = \emptyset$. For all $a \in S_1$, select $i(a)$ such that $a \notin S_{i(a)}$. Then the intersection of the at most $(r+1)$ sets S_1 and $S_{i(a)}$ is empty. □

For every base set size n and r, let $f(n,r)$ denote the maximum size of an r-superimposed family, and $g(n,r)$ denote the maximum size of an r-SUT family. In what follows, we give bounds on the *rate* of r-SUT families, which is

$$R_g(r) = \limsup_{n \to \infty} \frac{\lg g(n,r)}{n}.$$

Theorem 3.4: There exist constants $c_1, c_2 > 0$ such that

$$\frac{c_1}{r^2} \le R_g(r) \le \frac{c_2}{r}. \tag{1}$$

Proof of the Lower Bound: Clearly, if $\mathcal{F}$ is r-superimposed then it is r-SUT. Therefore,

$$g(n,r) \ge f(n,r) \ge 2^{c_1 n / r^2}$$

where the latter inequality can be found, say, in [3]. This gives the lower bound in (1). ♢

In order to prove the upper bound, we relate r-SUT to another property investigated in [21], [22].

Definition 3.5: (Alon, Fachini, Körner, [21]) A family $\mathcal{F}$ is r-locally thin if for all subsets $\mathcal{F}' \subseteq \mathcal{F}$ with $|\mathcal{F}'| = r$, there exists $x \in [n]$ such that

$$\sum_{A \in \mathcal{F}'} \{x \in A\} = 1$$

i.e., there exists an element x that appears in exactly one member of $\mathcal{F}'$.

We need the following strengthening of this definition.

Definition 3.6: A family $\mathcal{F}$ is $\le r$-locally thin if for all subsets $\mathcal{F}' \subseteq \mathcal{F}$ with $1 \le |\mathcal{F}'| \le r$, there exists such $x \in [n]$ that

$$\sum_{A \in \mathcal{F}'} \{x \in A\} = 1.$$

Lemma 3.7: If $\mathcal{F}$ is r-SUT then it is $\le (r+1)$-locally thin.

Proof: Contrary to the lemma, assume that there is a subset $\mathcal{F}' = \{A_1, \dots, A_k\}$, $1 \le k \le r+1$ for which $\sum_{i=1}^{k} \{x \in A_i\} \ne 1$ holds for all $x \in [n]$. For $i = 1, \dots, k$, let $\mathcal{F}_i = \mathcal{F}' - \{A_i\}$. Since every element is covered at least twice by the members of $\mathcal{F}'$

$$\bigcup_{A \in \mathcal{F}_1} A = \bigcup_{A \in \mathcal{F}_2} A = \cdots = \bigcup_{A \in \mathcal{F}_k} A, \quad \text{while} \quad \bigcap_{j=1}^{k} \mathcal{F}_j = \emptyset.$$

The existence of $\mathcal{F}_1, \dots, \mathcal{F}_k$ contradicts the r-SUT property. □

Let $h'(n,r)$, $h^*(n,r)$ be the maximum size of r-*locally thin*, $\le r$-*locally thin* families, respectively.

Corollary 3.8:

$$g(n,r) \le h^*(n, r+1) \le h'(n, r+1). \tag{2}$$

Proof: Here the first inequality comes from Lemma 3.7, while the second one follows directly from the definitions. □

Alon, Fachini, and Körner [21] proved the following theorem.

Theorem 3.9:

$$R_{h'}(r) < \frac{2}{r}, \qquad \text{for } r \text{ even}$$
$$R_{h'}(r) < \frac{c \log r}{r}, \qquad \text{for } r \text{ odd}, \ c \text{ is constant}. \tag{3}$$

Proof of the Upper Bound in Theorem 3.4: If r is odd, then $(r+1)$ is even. Hence, by (2) and (3)

$$R_g(r) \le R_{h^*}(r+1) \le R_{h'}(r+1) < \frac{2}{r+1}.$$

If r is even, then by the monotonicity of $h^*(n,r)$, (2), and (3)

$$R_g(r) \le R_{h^*}(r+1) \le R_{h^*}(r) \le R_{h'}(r) < \frac{2}{r}.$$

In either case, the upper bound holds in (1) with $c_2 = 2$. □

The following Lemma 3.10 allows for an alternative, self-contained proof of our upper bound on R_g, without using the (strong) bounds of Theorem 3.9. It gives a sufficient upper bound for $h^*(n,r)$ when r is even, which can then be employed with the monotonicity argument.

Lemma 3.10: Let r be even. If $\mathcal{F}$ is $\le r$-locally thin, then the modulo two sums of $(r/2)$-sets of characteristic vectors associated with members of $\mathcal{F}$ are all different.

Proof: For the sake of contradiction, assume that there are two collections $\mathcal{F}_1$ and $\mathcal{F}_2$ with the same modulo two sums. Consider the symmetric difference $\mathcal{F}' = \mathcal{F}_1 \triangle \mathcal{F}_2$. Clearly, it contains at most r sets, and every element in $\cup_{A \in \mathcal{F}'} A$ is covered at least twice (in fact, even times) by members of $\mathcal{F}'$. □

Corollary 3.11: If r is even, then $R_{h^*}(r) \le \frac{2}{r}$.

Proof: By Lemma 3.10, $\binom{h^*(n,r)}{r/2} \le 2^n$. □

IV. Disjointly r-Superimposed Codes

Another important case implicated in the multiple-access model of Section I is when the receiver must distinguish only between disjoint sets of active users. The following definition captures this notion.

Definition 4.1: A family $\mathcal{F} \subseteq 2^{[n]}$ is disjointly r-superimposed if

$$\bigcup_{i=1}^{k} A_i \ne \bigcup_{j=1}^{\ell} B_j \tag{4}$$

is implied by

$$\{A_1, A_2, \dots, A_k\} \cap \{B_1, B_2, \dots, B_\ell\} = \emptyset$$

for all $1 \le k, \ell \le r$; $A_1, A_2, \dots, A_k, B_1, B_2, \dots, B_\ell \in \mathcal{F}$.

Despite the seemingly slight difference between Definitions 4.1 and 3.1, the extremal properties of disjointly r-superimposed families and r-superimposed ones are completely different.

Let $h(n,r)$ be the maximum size of disjointly r-superimposed families.

Lemma 4.2: If $\mathcal{F}$ is r-superimposed then it is r-SUT. If $\mathcal{F}$ is r-SUT, then it is disjointly r-superimposed. Hence,

$$f(n,r) \le g(n,r) \le h(n,r).$$

Proof: The first part is already proved. The second part follows from the fact that if $\mathcal{F}$ is not disjointly r-superimposed, then there exist

$$\mathcal{A} = \{A_1, \dots, A_k\} \subseteq \mathcal{F} \quad \text{and} \quad \mathcal{B} = \{B_1, \dots, B_\ell\} \subseteq \mathcal{F}$$

such that $\cup_{i=1}^{k} A_i = \cup_{j=1}^{\ell} B_j$ while $\mathcal{A} \cap \mathcal{B} = \emptyset$. □

While we do not know if there is an exponential gap between r-superimposed and r-SUT families, the following theorem shows that there is such a gap between r-superimposed and disjointly r-superimposed ones.

Theorem 4.3: The rate of disjointly r-superimposed codes is bounded as

$$\frac{1}{2r} \le R_h(r) \le \left(\frac{1}{2} + o(1)\right) \frac{\lg r}{r}. \tag{5}$$

The key to the upper bound is the following observation.

Lemma 4.4: If $\mathcal{F}$ is disjointly r-superimposed then the vector sums of r-size sets of characteristic vectors associated with members of $\mathcal{F}$ are all different.

Proof: For the sake of contradiction, assume that there are two collections $\mathcal{F}_1, \mathcal{F}_2 \in \binom{\mathcal{F}}{r}$, with the same vector sums. Consider $\mathcal{F}_1' = \mathcal{F}_1 \setminus \mathcal{F}_2$ and $\mathcal{F}_2' = \mathcal{F}_2 \setminus \mathcal{F}_1$. Clearly, $|\mathcal{F}_1'|, |\mathcal{F}_2'| \le r$, and the vector sums of members of $\mathcal{F}_1'$ and $\mathcal{F}_2'$ are the same. But then

$$\cup_{A\in\mathcal{F}_1'} A = \cup_{B\in\mathcal{F}_2'} B$$

while $\mathcal{F}_1'$ and $\mathcal{F}_2'$ are disjoint, which is a contradiction. □

Now, in a vector sum $\boldsymbol{y} = (y^1, \ldots, y^n)$ of r binary vectors, $0 \le y^i \le r$ holds in every coordinate i. The number of possible vector sums is thus $(r+1)^n$, and therefore,

$$\binom{h(n,r)}{r} \le (r+1)^n$$

must hold. This gives an upper bound with a constant factor of 1 in (5). In order to obtain the factor of $\frac{1}{2}$, we use a second-moment method combined with a volume argument: we show that coordinates of almost all vectors in $\mathcal{F}$ deviate within $\sqrt{r}$ around the mean (instead of $r/2$, as above). In fact, we show that if a family $\mathcal{F}$ of subsets of $[n]$ has the property that for every choice of r sets, the sum of the corresponding characteristic vectors gives a different value, then the upper bound in (5) already holds. We prove Theorem 4.3 after Lemma 4.5 below.

For a set $\mathcal{A} \subseteq \{0,1\}^n$ of binary vectors of length n, $\boldsymbol{s}(\mathcal{A})$ stands for the sum of its elements

$$\boldsymbol{s}(\mathcal{A}) = \sum_{\boldsymbol{x}\in\mathcal{A}} \boldsymbol{x}.$$

Lemma 4.5: Let $\mathcal{F}$ be a set of binary vectors of length n, and let $T = |\mathcal{F}|$. Let $\boldsymbol{c} = T^{-1}\sum_{\boldsymbol{v}\in\mathcal{F}} \boldsymbol{v}$ be the average vector of the set. For every integer $1 \le r \le T$, the inequality

$$\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left\|\boldsymbol{s}(\mathcal{A}) - r\boldsymbol{c}\right\|^2 \le nr\binom{T}{r}$$

holds, where $\|\cdot\|$ is the Euclidean norm.

Proof: By definition of the norm

$$\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left\|\boldsymbol{s}(\mathcal{A}) - r\boldsymbol{c}\right\|^2 = \sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \|\boldsymbol{s}(\mathcal{A})\|^2 + \sum_{\binom{\mathcal{F}}{r}} r^2\|\boldsymbol{c}\|^2 - \sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} 2r\boldsymbol{c}\boldsymbol{s}(\mathcal{A}). \quad (6)$$

Clearly, the second term in (6) gives $\binom{T}{r}r^2\|\boldsymbol{c}\|^2$. The third term is

$$\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} 2r\boldsymbol{c}\boldsymbol{s}(\mathcal{A}) = 2r\boldsymbol{c}\binom{T-1}{r-1}\sum_{\boldsymbol{v}\in\mathcal{F}}\boldsymbol{v} = 2\binom{T}{r}r^2\|\boldsymbol{c}\|^2$$

since in the sum $\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \boldsymbol{s}(\mathcal{A})$ every vector $\boldsymbol{v}\in\mathcal{F}$ appears with multiplicity $\binom{T-1}{r-1}$, which is the number of distinct r-sets in which a given vector $\boldsymbol{v}$ is contained. The first term of (6) can be bounded as follows:

$$\begin{aligned}\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \|\boldsymbol{s}(\mathcal{A})\|^2 &= \sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left\|\sum_{\boldsymbol{v}\in\mathcal{A}}\boldsymbol{v}\right\|^2 \\ &\le \sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left(nr + 2\sum_{\substack{1\le i<j\le r\\ \boldsymbol{v}_i,\boldsymbol{v}_j\in\mathcal{A}}} \boldsymbol{v}_i\boldsymbol{v}_j\right) \\ &= nr\binom{T}{r} + 2\binom{T-2}{r-2}\sum_{\substack{1\le i<j\le T\\ \boldsymbol{v}_i,\boldsymbol{v}_j\in\mathcal{F}}} \boldsymbol{v}_i\boldsymbol{v}_j \\ &= nr\binom{T}{r} + 2\binom{T-2}{r-2}\sum_{\substack{1\le i<j\le T\\ \boldsymbol{v}_i,\boldsymbol{v}_j\in\mathcal{F}}} \boldsymbol{v}_i\boldsymbol{v}_j \\ &\quad + \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 \\ &= nr\binom{T}{r} + \binom{T-2}{r-2}\left\|\sum_{\boldsymbol{v}\in\mathcal{F}}\boldsymbol{v}\right\|^2 - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 \\ &= nr\binom{T}{r} + \binom{T-2}{r-2}T^2\|\boldsymbol{c}\|^2 - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2.\end{aligned}$$

For the inequality, we used that the norm square of every vector is at most n, as every vector is binary. Subsequently, we used that every pair of vectors appears together in exactly $\binom{T-2}{r-2}$ sets of size r, and thus, every product $\boldsymbol{v}_i\boldsymbol{v}_j$ occurs that many times.

Returning to (6), by the above computation we get

$$\begin{aligned}\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left\|\boldsymbol{s}(\mathcal{A}) - r\boldsymbol{c}\right\|^2 &\le nr\binom{T}{r} + \binom{T-2}{r-2}T^2\|\boldsymbol{c}\|^2 - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 \\ &\quad - \binom{T}{r}r^2\|\boldsymbol{c}\|^2 \\ &= nr\binom{T}{r} + \binom{T-2}{r-2}T^2\|\boldsymbol{c}\|^2 - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 \\ &\quad - \binom{T-2}{r-2}r^2\frac{T(T-1)}{r(r-1)}\|\boldsymbol{c}\|^2.\end{aligned}$$

From $r \le T$ follows that

$$-r^2\frac{T(T-1)}{r(r-1)} \le -T^2.$$

Therefore,

$$\begin{aligned}\sum_{\mathcal{A}\in\binom{\mathcal{F}}{r}} \left\|\boldsymbol{s}(\mathcal{A}) - r\boldsymbol{c}\right\|^2 &\le nr\binom{T}{r} - \binom{T-2}{r-2}\sum_{\boldsymbol{v}\in\mathcal{F}}\|\boldsymbol{v}\|^2 \\ &\quad + \binom{T-2}{r-2}T^2\|\boldsymbol{c}\|^2 - \binom{T-2}{r-2}T^2\|\boldsymbol{c}\|^2\end{aligned}$$

which implies to the desired result. □

Proof of Theorem 4.3: First we prove the upper bound. Take an arbitrary set $\mathcal{F} \subseteq \{0,1\}^n$ of binary vectors of length n, such that the vector sums are different for all choices of r vectors. (By Lemma 4.4, the set of characteristic vectors for a disjointly r-superimposed family fulfills this condition.) Let $T = |\mathcal{F}|$. As in Lemma 4.5, define the average vector $\boldsymbol{c} = T^{-1}\sum_{\boldsymbol{v}\in\mathcal{F}}\boldsymbol{v}$. Let $\mathcal{A} \subseteq \mathcal{F}$ be a random subset of size r, chosen with uniform probability. Consider the random variable $\xi = \|\boldsymbol{s}(\mathcal{A}) - r\boldsymbol{c}\|$, the distance of $\boldsymbol{s}(\mathcal{A})$ from its mean. By Lemma 4.5 and Jensen's inequality [23], the expected distance $\mathbb{E}\xi \le \sqrt{nr}$. By Markov's inequality [23]

$$\mathbb{P}\{\xi \ge \lambda^{-1}\sqrt{nr})\} \le \lambda$$

for all $0 < \lambda < 1$. This means that for any constant $0 < \lambda < 1$, at least the $(1-\lambda)$ fraction of all sums for r-size subsets of $\mathcal{F}$ lie within the n-dimensional ball B of radius $\lambda^{-1}\sqrt{nr}$ centered at the point $r\boldsymbol{c}$. Therefore, the number of integer lattice points in B is an upper bound for $(1-\lambda)\binom{T}{r}$. Consider a larger ball B' with radius $(\sqrt{nr}/\lambda + \sqrt{n}/2)$ centered at $r\boldsymbol{c}$. Its volume bounds the number of lattice points in B from above. To see this, draw an n-dimensional unit cube centered at each lattice point in B. All the cubes are within B', and to each integer lattice point a unit volume is associated. Using the well-known formula for the volume of an n-dimensional ball (e.g., [24])

$$(1-\lambda)\binom{T}{r} \le \frac{\pi^{n/2}\left(\lambda^{-1}\sqrt{nr}+\frac{1}{2}\sqrt{n}\right)^n}{\Gamma(1+n/2)}$$

where $\Gamma(x)$ is the complete gamma function. An application of Stirling's approximation [23] to bound $\Gamma(1+n/2)$ leads to

$$\frac{\lg T}{n} \le \frac{\lg r}{2r} + \Theta\left(\frac{1}{r}\right) + \frac{o(n)}{n}$$

which is tantamount to the upper bound of (5).

We prove the lower bound in (5) with a probabilistic argument. (This proof was also observed by László Györfi.)

Let $\mathcal{F}$ be a randomly constructed family of size T, where T will be specified later. Every set $A_i \in \mathcal{F}$ is constructed randomly so that $x \in A_i$ with probability $(1-2^{-1/r})$ for all x independently. We prove that $\mathcal{F}$ is disjointly r-superimposed with nonzero probability for some $T = 2^{\Theta(n/r)}$. Let $\mathcal{A}, \mathcal{B} \subseteq \mathcal{F}$ be two disjoint sets: $\mathcal{A} = \{A_1, \dots, A_k\}$ and $\mathcal{B} = \{B_1, \dots, B_\ell\}$, where $\mathcal{A} \cap \mathcal{B} = \emptyset$ and $1 \le k, \ell \le r$. Define $A = \cup_{i=1}^{k} A_i$ and $B = \cup_{j=1}^{\ell} B_j$. Equation (4) is violated if for all $x \in [n]$, either $x \in A \cap B$ or $x \notin A \cup B$. Since all A_i and B_j are independent

$$\begin{aligned} p(k,\ell) &= \mathbb{P}\bigcap_{x\in[n]}\left(\{x \notin A; x \notin B\} \cup \{x \in A; x \in B\}\right) \\ &= \left(p^{k+\ell} + (1-p^k)(1-p^\ell)\right)^n \qquad (7) \end{aligned}$$

where $p = 2^{-1/r}$. The expected number of disjoint set pairs that violate (4) is thus

$$N = \sum_{k=1}^{r}\sum_{\ell=0}^{k-1}\binom{T}{k}\binom{T-k}{\ell}p(k,\ell) + \sum_{k=1}^{r}\binom{\binom{T}{k}}{2}p(k,k). \qquad (8)$$

By the choice of p and the fact that $k, \ell \le r$, $(1-p^k) \le p^k$ and $(1-p^\ell) \le p^\ell$. Consequently, the right-hand side of (7) is bounded from above as $p(k,\ell) \le \min\{p^{nk}, p^{n\ell}\}$. Hence, the right-hand side of (8) is bounded from above as

$$N \le \sum_{k=1}^{r}\sum_{\ell=0}^{k-1}\binom{T}{k}\binom{T-k}{\ell}p^{nk} + \sum_{k=1}^{r}\binom{\binom{T}{k}}{2}p^{nk}. \qquad (9)$$

Now, for $T \ge 2r^2 + r - 1$

$$\begin{aligned} \sum_{\ell=0}^{k-1}\binom{T-k}{\ell} &\le \sum_{\ell=0}^{k-1}\binom{T}{\ell} \le k\binom{T}{k-1} \\ &= \frac{k^2}{T-k+1}\binom{T}{k} \le \frac{1}{2}\binom{T}{k} \end{aligned}$$

and

$$\binom{\binom{T}{k}}{2} < \left(\binom{T}{k}\right)^2 \Big/ 2.$$

Subsequently, (9) is bounded by

$$N \le \sum_{k=1}^{r}\left(\binom{T}{k}\right)^2 p^{nk}. \qquad (10)$$

In (10), the largest term is the one for $k = 1$ if $T \le (k+1)p^{-n/2} + k$ for all k, i.e., if

$$T \le 1 + 2p^{-n/2} = 1 + 2\cdot 2^{\frac{n}{2r}}. \qquad (11)$$

Then by (10)

$$N \le rT^2p^n$$

and thus $N < 1$ if

$$T < \frac{p^{-n/2}}{\sqrt{r}} = \frac{2^{\frac{n}{2r}}}{\sqrt{r}}. \qquad (12)$$

Between (11) and (12), (12) is more restrictive for all n and r. As a consequence, there exists a disjointly r-superimposed family of size $T = \left\lceil r^{-1/2}2^{\frac{n}{2r}}\right\rceil - 1$, which implies the lower bound of (5). □

V. Open Problems

We conclude by posing the following open problems.

Problem 5.1: It is known that

$$\frac{c_1}{r^2} \le R_f(r) \le \frac{c_2 \lg r}{r^2}.$$

Try to diminish the gap between the two bounds.

Problem 5.2: We show in this correspondence that

$$\frac{c_1}{r^2} \le R_g(r) \le \frac{c_2}{r}.$$

Try to diminish the gap between the two bounds.

Problem 5.3: We show in this correspondence that

$$\frac{1}{2r} \le R_h(r) \le \left(\frac{1}{2} + o(1)\right)\frac{\lg r}{r}.$$

Try to diminish the gap between the two bounds.

Problem 5.4: Do r-SUT and r-superimposed families differ significantly, i.e., do the functions $R_g(r)$ and $R_f(r)$ differ in magnitude?

Remark: In the course of submitting this correspondence we learned that Noga Alon and Vera Asodi showed that $R_g(r) = \Omega(1/r)$ which answers Problems 5.2 and 5.4.

Acknowledgment

The authors wish to thank Sándor Győri for commenting on an earlier version of the manuscript. They would also like to thank Noga Alon and László Györfi for fruitful discussions.

References

[1] A. G. D'yachkov and V. V. Rykov, "Bounds on the length of disjunctive codes," *Probl. Pered. Inf.*, vol. 18, no. 3, pp. 7–13, 1982.

[2] P. Erdős, P. Frankl, and Z. Füredi, "Families of finite sets in which no set is covered by the union of r others," *Israel J. Math.*, vol. 51, pp. 79–89, 1985.

[3] F. K. Hwang and V. T. Sós, "Non-adaptive hypergeometric group testing," *Stud. Sci. Math. Hungar.*, vol. 22, pp. 257–263, 1987.

[4] Z. Füredi, "A note on r-cover-free families," *J. Combin. Theory Ser. A*, vol. 73, pp. 172–173, 1996.

[5] W. H. Kautz and R. C. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inf. Theory*, vol. IT-10, no. 4, pp. 363–377, Oct. 1964.

[6] M. Ruszinkó, "On the upper bound of the size of the r-cover-free families," *J. Combin. Theory Ser. A*, vol. 66, pp. 302–310, 1994.

[7] C. C. Colbourn, A. C. H. Ling, and M. Tompa, "Construction of optimal quality control for oligo arrays," *Bioinformatics*, vol. 18, no. 4, pp. 529–535, 2002.

[8] R. Beigel, N. Alon, S. Kasif, M. S. Apaydin, and L. Fortnow, "An optimal procedure for gap closing in whole genome shotgun sequencing," in *Proc. 5th Annu. Int. Conf. Computational Biology (RECOMB)*, Montreal, QC, Canada, Apr. 22–25, 2001, pp. 22–30.

[9] D. J. Balding, W. J. Bruno, E. Knill, and D. C. Torney, "A comparative survey of nonadaptive pooling designs," in *Genetic Mapping and DNA Sequencing*, T. Speed and M. S. Waterman, Eds. New York: Springer-Verlag, 1996, vol. 81, IMA volumes in mathematics and its applications, pp. 133–154.

[10] E. D. Green, "Strategies for the systematic sequencing of complex genomes," *Nat. Rev. Genet.*, vol. 2, pp. 573–583, 2001.

[11] "International human genome sequencing consortium, “Initial sequencing and analysis of the human genome”," *Nature*, vol. 609, no. 6822, pp. 860–921, 2001.

[12] E. Barillot, B. Lacroix, and D. Cohen, “Theoretical analysis of library screening using an $\boldsymbol{n}$-dimensional strategy,” *Nucl. Acids Res.*, vol. 19, pp. 6241–6247, 1991.

[13] W. J. Bruno, E. Knill, D. J. Balding, D. C. Bruce, N. A. Doggett, W. W. Sawhill, R. L. Stallings, C. C. Whittaker, and D. C. Torney, “Efficient pooling designs for library screening,” *Genomics*, vol. 26, pp. 21–30, 1995.

[14] M. A. Chateauneuf, C. J. Colbourn, D. L. Kreher, E. R. Lamken, and D. C. Torney, “Pooling, lattice square, and Union Jack designs,” *Ann. Combin.*, vol. 3, pp. 27–35, 1999.

[15] A. G. D’yachkov, A. J. Macula Jr., and V. V. Rykov, “New constructions of superimposed codes,” *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 284–290, Jan. 2000.

[16] A. J. Macula, “Probabilistic nonadaptive group testing in the presence of errors and DNA library screening,” *Ann. Combin.*, vol. 3, pp. 61–69, 1999.

[17] H. Q. Ngo and D.-Z. Du, “New constructions of nonadaptive and error-tolerance pooling designs,” *Discr. Math.*, vol. 243, pp. 161–170, 2002.

[18] W.-W. Cai, R. Chen, R. A. Gibbs, and A. Bradley, “A clone-array pooled strategy for sequencing large genomes,” *Genome Res.*, vol. 11, pp. 1619–1623, 2001.

[19] M. Csűrös and A. Milosavljevic, “Pooled genomic indexing (PGI): Analysis and design of experiments,” *J. Comput. Biol.*, vol. 11, no. 5, pp. 1001–1021, 2004.

[20] E. S. Lander and M. S. Waterman, “Genomic mapping by fingerprinting random clones: A mathematical analysis,” *Genomics*, vol. 2, pp. 231–239, 1988.

[21] N. Alon, E. Fachini, and J. Körner, “Locally thin set families,” *Comb., Prob. Comput.*, vol. 9, pp. 481–488, 2000.

[22] Z. Füredi, A. Gyárfás, and M. Ruszinkó, “On the maximum size of (p, Q)-free families,” *Discr. Math.*, vol. 257, pp. 385–403, 2002.

[23] W. Feller, *An Introduction to Probability Theory and its Applications*. New York: Wiley, 1966.

[24] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices, and Groups*, 2nd ed. New York: Springer-Verlag, 1993.

A New Lower Bound for Multiple Hypothesis Testing

Lucien Birgé

***Abstract*—The purpose of this correspondence is to give a new, easily tractable, and sharp lower bound for the maximal error in multiple hypothesis testing with an application to nonasymptotic lower bounds for the minimax risk of estimators.**

***Index Terms*—Fano’s lemma, minimax risk, multiple hypothesis testing.**

I. Introduction

A classical benchmark for the quality of an estimator of some unknown parameter θ belonging to a set Θ is the minimax risk which is defined as follows. Assume we want to estimate the parameter θ belonging to some metric space (Θ, d) from one observation $\boldsymbol{X}$ with unknown distribution $P_\theta, \theta \in \Theta$ and we consider a loss function of the form $w[d(\theta, \theta')]$ where w is nonnegative and nondecreasing. The *minimax risk* $R_M(\Theta)$ over Θ is then given by $R_M(\Theta) = \inf_{\hat{\theta}} R(\hat{\theta}, \Theta)$ with

$$R(\hat{\theta}, \Theta) = \sup_{\theta \in \Theta} \mathbb{E}_\theta[w(d(\theta, \hat{\theta}(\boldsymbol{X})))]$$

where the infimum is over all (possibly randomized) estimators $\hat{\theta}(\boldsymbol{X})$ with values in Θ.

Since it is typically impossible to compute $R_M(\Theta)$ exactly, one merely tries to bound it from below as accurately as possible. For this, a quite classical way is to introduce a finite subset Θ' of Θ such that $d(\theta, \theta') \geq \eta$ for all pairs $\theta \neq \theta'$ belonging to Θ' and use the fact that

$$R_M(\Theta) \geq R_M(\Theta') \geq w\left(\frac{\eta}{2}\right) \inf_T \sup_{\theta \in \Theta'} \mathbb{P}_\theta[T \neq \theta] \tag{1.1}$$

where T denotes an arbitrary estimator with values in Θ'. The proof is straightforward and can be found, for instance, in Yu [1]. It can be extended to the case when d is not a genuine distance but satisfies some sort of a triangular inequality as in Yang and Barron [2, pp. 1570–1571]. With this approach, given the subset Θ' with cardinality $N + 1$, bounding $R_M(\Theta)$ from below reduces to solving the following problem.

Multiple hypothesis testing problem Given a family $\{P_0, \ldots, P_N\}$ of probability measures on some measurable set $(E, \mathcal{E})$ and a random variable $\boldsymbol{X}$ with an unknown distribution in the family, find a lower bound for the maximum probability of error

$$p_M = \inf_T \sup_{0 \leq i \leq N} \mathbb{P}_i[T \neq i]$$

where T denotes an arbitrary (possibly randomized) estimator based on $\boldsymbol{X}$ with values in the set of indices $\{0, \ldots, N\}$ and $\mathbb{P}_i$ the probability that gives $\boldsymbol{X}$ the distribution P_i.

Equivalently, p_M can be viewed as the maximal error for decoding the output from a noisy channel when P_i is the output distribution corresponding to input i.

One way of solving the above problem, that has been extensively used in the nonparametric statistics literature since its introduction in the field by Ibragimov and Has’minskii, is to use Fano’s inequality (see Fano [3]), as stated in Gallager [4, p. 77, Theorem 4.3.1], and Cover and Thomas [5, p. 39]. One first observes that

$$p_M \geq p_e = \inf_T (N+1)^{-1} \sum_{i=0}^{N} \mathbb{P}_i[T \neq i]$$

and then applies Fano’s inequality which provides a lower bound for p_e based on the Kullback–Leibler divergence (KL divergence, for short) defined by

$$K(P, Q) = \begin{cases} \int \log(dP/dQ)dP, & \text{if } P \ll Q \\ +\infty, & \text{otherwise.} \end{cases}$$

This leads to the following lower bound for p_M which can be found, for instance, in the book by Ibragimov and Has’minskii [6, p. 25]: $p_M \geq 1 - (\overline{K} + \log 2)/(\log N)$ with

$$\overline{K} = \frac{1}{N+1} \sum_{i=0}^{N} K\left(P_i, \frac{1}{N+1} \sum_{i=0}^{N} P_i\right). \tag{1.2}$$

Various versions of this inequality have been used in the nonparametric statistics literature under the name of “Fano’s lemma” or “Fano’s method” in order to derive lower bounds for the minimax risk. Let us mention first the seminal works of Ibragimov and Has’minskii [7], [8] and [6] or Has’minskii [9], then, among many other references, Birgé [10] and [11], Yang and Barron [2], and Nemirovski [12]. For a nice presentation of the method and related results, see [1] and the references therein.

Related bounds, with a different emphasis, can be found in the sequential analysis literature. There, the problem is to get lower bounds for the expected number of tries in a sequential multiple hypothesis problem with fixed errors $a_{i,j} = \mathbb{P}_i[T = j]$ for $j \neq i$. This problem can be viewed as a dual of ours. Its solution has been found by Wald [13] for binary tests $(N = 1)$ and has been extended to multiple testing by Simons [14]. These bounds easily translate to the nonsequential

Manuscript received April 10, 2001; revised October 28, 2004.

The author is with the UMR CNRS 7599 “Laboratoire de Probabilités et Modèles Aléatoires,” Université Paris VI, F-75252 Paris Cedex 05, France (e-mail: lb@ccr.jussieu.fr).

Communicated by P. Narayan, Associate Editor for Shannon Theory.

Digital Object Identifier 10.1109/TIT.2005.844101