Efficient reconstruction of sequences

Vladimir I. Levenshtein

Abstract— **In this paper we introduce and solve some new problems of effective reconstruction of an unknown sequence of a fixed length with the help of its patterns which are distorted by errors of a given type. These patterns can be considered as output sequences for multiple transmission of this sequence over a combinatorial channel with a maximum permissible number of single errors or over a discrete probabilistic channel. The efficiency of recognition is understood in the sense of minimizing the number N of erroneous patterns which are sufficient to reconstruct exactly any unknown sequence or reconstruct it with a preset accuracy and/or with a given probability. This also includes the existence of simple reconstruction algorithms. Complete solutions for combinatorial channels with some types of errors of essential interest in coding theory, namely: substitutions, transpositions, deletions and insertions of symbols are given. For these cases simple reconstruction algorithms based on majority and threshold principles and their non-trivial combination are found. In general, for combinatorial channels the considered problem is reduced to a new problem of reconstructing a vertex of an arbitrary graph with the help of the minimum number of vertices in its metrical ball of a given radius. A certain sufficient condition for solution of this problem is presented. For a discrete memoryless channel asymptotic behavior of the minimum number of repeated transmissions which are sufficient to reconstruct any sequence of length n within the Hamming distance d with error probability ε is found when d/n and ε tends to 0 as $n \to \infty$. A similar result for the continuous channel with discrete time and additive Gaussian noise is also obtained.**

Index Terms - **Reconstruction, sequences, optimization, algorithms, graphs, the error metric, combinatorial and probabilistic channels, repeated transmission, reducible reconstructors, asymptotic behavior**

I. Introduction

Traditional problems of theory of information transmission consist in efficient transmission of messages of a set V over noisy channels which are described by combinatorial or probabilistic conditions. For solution of these problems a *coding* elements of V is used to introduce a redundancy to messages so that the distance (in a certain metric) between the encoded messages would be sufficiently large and allows one to correct errors at the output of channels. The efficiency of transmission is characterized by amount of redundancy and by complexity of decoding algorithm. One of the known methods used to introduce redundancy is multiple transmission of a message over the same channel. However, this method needs too large redundancy and is not efficient as compared to other methods. Nevertheless, in various fields of natural science, such as informatics, molecular biology, chemistry, there is an important problem when the repeated transmission is in fact a unique method of recovering information. This problem consists of reconstruction of an unknown sequence (or message) $x = (x_1, ..., x_n)$ from a given set V (without a coding of elements which can be close one to another). We can assume that the components of x belong to the alphabet $F_q = \{0, 1, ..., q-1\}$, $q \geq 2$, and hence $V \subseteq F_q^n$. The sequence x can be lost and the only opportunity for its reconstruction is the knowledge of a sufficiently large number N of its patterns distorted by errors.

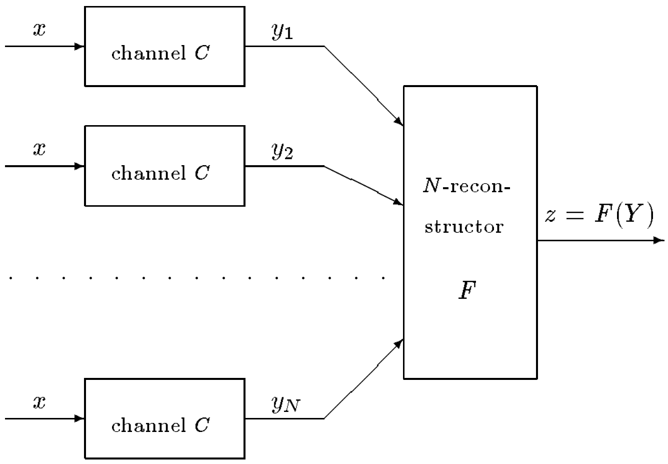

Fig. 1. Schematic diagram of reconstructing a sequence x

For the precise formulation of the problem we can define a combinatorial or probabilistic channel C and assume that these N patterns $y_1, ..., y_N$ are obtained by multiple transmission of x over the same channel C (see Fig. 1). For a chosen type of single errors (for instance, substitutions, transpositions or deletions of symbols), a combinatorial channel C is defined by the maximum number t of single errors which can occur during transmission of any input sequence of length n over the channel. As probabilistic channel C we consider an ordinary discrete noisy channel without memory [19], [7]. For restoring a sequence $x \in V$ in the both cases we use an N-reconstructor F which maps the matrix Y formed by the columns $y_1, ..., y_N$ to V. For combinatorial and probabilistic channels with substitution errors elements of each row of Y are obtained from the same symbol. We call $F(Y)$ an *exact reconstruction* of x if $F(Y) = x$, and a *reconstruction of x within distance d* if $d_H(F(Y), x) \leq d$ where $d_H(\cdot, \cdot)$ is the Hamming distance. (The case $d = 0$ corresponds to exact reconstruction.) In the case of a probabilistic channel C for any $x \in V$ the probability of the event $d_H(F(Y), x) > d$ characterizes an

V.I. Levenshtein is with the Keldysh Institute for Applied Mathematics, Russian Academy of Science, 125047 Moscow, Russia. His research was supported by the Russian Foundation for Basic Research (grant 98-01-00146). The material in this work was presented in part at the IX All Union Conference on Coding Theory and Information Transmission, Odessa, Russia, 1988, at the Fifth International Workshop on Algebraic and Combinatorial Coding Theory, Sozopol, Bulgaria, June 1-7, 1996, at the Second Sibirean Congress on Applied and Industrial Mathematics, Novosibirsk, Russia, July 25-29, 1996, and at the IEEE International Symposium on Information Theory, Ulm, Germany, July 29 - July 4, 1997.

error probability of reconstruction of x within distance d. In the case of a combinatorial channel C we should certainly assume that all patterns $y_1, ..., y_N$ are *different*, since otherwise exact reconstruction of x is not possible, for instance, when they coincide and differ from x. This is not essential in the probabilistic case. Moreover, we shall see that even if the probability of errorless transmission of x over C equals zero, the error probability of exact reconstruction of x can be made as small as one wishes with the help of an N-reconstructor for sufficiently large N.

A natural measure of efficiency of a solution of the combinatorial problem under consideration is the minimum number $N = N(V,t) + 1$ such that there exists an N-reconstructor which exactly reconstructs any $x \in V$ with the help of any N of its different erroneous patterns (if they exist). It is also significant to find a simple realization of this mapping (reconstruction algorithm). Analogously, for a discrete probabilistic memoryless channel it is important to find the minimum number $N = N_C(n, d, \varepsilon)$ such that there exists an N-reconstructor whose error probability of reconstruction of any $x \in V$ within distance d does not exceed a given ε $(0 < \varepsilon < 1/2)$.

Now we briefly describe the main results of the paper. In Section II we define combinatorial channels which are useful to describe combinatorial problems of efficient reconstruction of sequences and give solutions of these problems in the case of single errors of the essential interest in coding theory. First we consider the combinatorial problem of reconstructing an unknown sequence $x = (x_1, ..., x_n) \in F_q^n$ with the help of N different sequences $y_1, y_2, ..., y_N \in F_q^n$ each of which differs from x in at most t components (i.e. is obtained with the help of substitutions at most t symbols). What is the minimum number N which is sufficient to exactly reconstruct any $x \in F_q^n$? Does there exist a simple procedure for such a reconstruction? For example, is the following matrix Y

$$\begin{matrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{matrix}$$

from eleven different sequences (written as columns) sufficient to reconstruct $x \in F_2^5$ when $t = 2$? For any n, q and t we show that this minimum number N equals $N(F_q^n, t) + 1$ where

$$N(F_q^n, t) = q \sum_{i=0}^{t-1} \binom{n-1}{i} (q-1)^i$$

and prove that in each i-th row, $i = 1, 2, .., n$, of the matrix Y formed by the columns $y_1, y_2, ..., y_N$, one letter of F_q occurs more often than others and equals to the component x_i of the unknown $x = (x_1, ..., x_n)$. Thus, in this case the majority algorithm can be applied to all rows of Y in order to reconstruct x. In particular, for $t = 2$ eleven sequences above are sufficient to reconstruct $x \in F_2^5$, and $x = (0, 0, 0, 0, 1)$. Note that if we remove the last column of Y, another solution $x = (1, 0, 0, 0, 1)$ would be also suitable.

The next combinatorial problem is to reconstruct an unknown sequence $x = (x_1, ..., x_n) \in F_2^n$ using the minimum number N of different sequences $y_1, y_2, ..., y_N \in F_2^n$ each of which can be obtained from x with the help of at most t transpositions of two components. Since these transpositions do not change the weight of the binary vector x, we can assume that $x \in J_w^n$ where $0 < w < n$ and J_w^n is a subset of F_2^n consisting of all vectors of weight w. As an example we consider the following question. Can we uniquely restore $x \in J_2^5$ by the matrix Y

$$\begin{matrix} 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \end{matrix}$$

from six different columns obtained from x with the help at most one transposition of two symbols? We show that this minimum number N equals $N^\circ(J_w^n, t) + 1$ where

$$N^\circ(J_w^n, t) = n \sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i} \frac{1}{i+1}.$$

We also prove that for reconstruction of $x = (x_1, ..., x_n) \in J_w^n$ by the matrix Y, whose columns are formed by its $N = N^\circ(J_w^n, t) + 1$ erroneous patterns, the following threshold algorithm can be applied : $x_i = 1$ if the number of ones in the i-th row of Y is greater than $(N-1)w/n$ and $x_i = 0$ otherwise $(i = 1, ..., n)$. In particular, $N = n + 1$ when $t = 1$ and the six sequences are sufficient to uniquely restore $x \in J_2^5$, and we have $x = (1, 0, 0, 1, 0)$. If we remove the last column, another solution $x = (0, 0, 0, 1, 1)$ would be suitable as well. We verify that $N = n + 1$ different patterns are also sufficient for reconstruction of any $x \in F_q^n$ when $q \geq 3$ and $t = 1$. However, this needs a generalized version of the threshold algorithm which will be described in Section II. We also give solutions to similar problems that allow asymmetric substitutions $0 \to 1$ or $1 \to 0$. In these cases, $x \in J_w^n$ can be efficiently reconstructed by applying Boolean functions, respectively disjunction and conjunction in N variables, to rows of Y.

A more complex combinatorial problem is connected with reconstruction of an arbitrary $x = (x_1, ..., x_n) \in F_q^n$ with the help of different sequences $y_1, y_2, ..., y_N \in F_q^{n-t}$ every of which is obtained from x by deletions of exactly t symbols and hence is a subsequence $(x_{i_1}, ..., x_{i_{n-t}})$ of x, $1 \leq i_1 < \cdots < i_{n-t} \leq n$. If $N^-(F_q^n, t)$ denotes the maximum size of the set of common subsequences of length $n - t$ of two different sequences x, $z \in F_q^n$, then $N = N^-(F_q^n, t) + 1$ is the minimum number such that any $x \in F_q^n$ can be exactly reconstructed using N of its different subsequences of length $n - t$ (if they exist). We find $N^-(F_q^n, t)$ and prove that

$$N^-(F_q^n, t) \leq q \sum_{i=0}^{t-1} \binom{n-t-1}{i} (q-1)^i \qquad (1)$$

with equality for $q = 2$. In particular, $N^-(F_2^7, 2) = 10$ and hence we can uniquely define a sequence $x \in F_2^7$ under condition that the eleven columns of the following matrix:

$$\begin{matrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \end{matrix}$$

are subsequences of length 5 of x. An algorithm for reconstruction of $x \in F_q^n$ with the help of $N = N^-(F_q^n, t) + 1$ of its different subsequences of length $n - t$ is based on an interesting combination of majority and threshold principles. This algorithm is described and illustrated by an example in Section II. We also find the maximum number $N^+(F_q^n, t)$ of common supersequences of length $n + t$ of two different sequences from F_q^n and present an algorithm for reconstruction of an arbitrary $x \in F_q^n$ by any of its $N = N^+(F_q^n, t) + 1$ different supersequences of length $n + t$ (they are obtained from x by insertions of t symbols of F_q).

The considered combinatorial problems show that in fact we deal with the same problem for different metrics on F_q^n. In this connection we advance in Section III a graph-theoretical approach to the problem of reconstructing an unknown sequence using the minimum number of its patterns distorted by errors of a given type and restricted multiplicity. This problem is reduced to a new problem of reconstructing a vertex of an arbitrary graph with the help of the minimum number of vertices in its metrical ball of a given radius. We consider a finite set V of *messages* and a set H of one-to-one, in general, partial mappings $V \to V$ (called *single errors*) which have the following property: if $x, y \in V$, $h \in H$, and $h(x) = y$, then there exists $g \in H$ such that $x = g(y)$. We define a graph $G = G(V, E)$ with the set V of vertices and set $E \subseteq V \times V$ of edges as follows: $(x, y) \in E$ if and only if $x \neq y$ and there exists $h \in H$ such that $h(x) = y$. One can show that the path distance $\rho(x, y)$ between vertices x and y of the graph G is equal to the minimum number of single errors translating x to y. This construction is applicable to many sorts of single errors of essential interest in coding theory such as substitutions, transpositions, bursts, deletions and insertions of symbols, and arithmetic errors. Moreover, we shall see that an arbitrary graph G of maximal degree r can be treated as a graph whose path distance is defined by a set H of r single errors. For an arbitrary graph G of diameter s and any integers t, d, $0 \leq t, d \leq s$, denote by $N(t, d)$ the maximum number of vertices in the intersection of metric balls of radius t around vertices x and y such that $d(x, y) \geq d$. The number $N(t, 1) + 1$ equals the minimum number of vertices in the metric ball of radius t around any vertex x sufficient for its exact reconstruction. The property of *monotony on intersections* is introduced and a sufficient condition so that a graph has this property is found. This property is held by some graphs and allows us to easily calculate the values $N(t, d)$.

It is significant to note that $N(t, d) + 1$ can be treated as the minimum number of vertices in the metric ball of radius t sufficient for *exact* reconstruction of a vertex of a code $C \subseteq V$ with the minimal distance d. This allows us to consider and solve some new problems of coding theory when the minimum distance of a code does not allow one to correct errors of a given multiplicity. In particular, for the binary Hamming graph, $N(t, 2t-1) = \binom{2t}{t}$ and this means that the minimum number of erroneous patterns sufficient for reconstruction of any word of a $(t-1)$-error-correcting code of length n in the combinatorial channel with at most t substitutions of symbols is equal to $\binom{2t}{t} + 1$ independent of length of the code.

The problem of efficient reconstruction of an unknown sequence at the output of a discrete probabilistic channel C without memory is considered in Section IV. This consists of finding the minimum number $N_C(n, d, \varepsilon)$ of repeated transmissions which are sufficient for reconstruction of any sequence of length n within the Hamming distance d with error probability ε. (The case $d = 0$ corresponds to exact reconstruction.) To estimate $N_C(n, d, \varepsilon)$ we introduce and study an optimal N-reconstructor and an optimal reducible N-reconstructor. Reducible N-reconstructors have a remarkable property: N-tuple transmission of a message over a discrete memoryless channel C with using a reducible N-reconstructor (see Fig. 1) is equivalent to *unitary* transmission of this message over a certain discrete memoryless channel C_N which has an "improved" transition matrix. A channel C is referred to as *non-degenerate* if its transition matrix does not have two identical rows (otherwise there exists a sequence which cannot be reconstructed with error probability $\varepsilon < 1/2$) and contains a column with at least two nonzero probabilities (otherwise any output sequence allows us to reconstruct exactly the input sequence). For a channel C we consider a constant $\alpha(C)$ which was in fact introduced in [21] for finding the zero rate exponent and note that $0 < \alpha(C) < 1$ if and only if C is non-degenerate. The main result can be formulated as follows. Let $\varepsilon = \varepsilon(n)$ and $d = d(n)$ be functions such that $\varepsilon \to 0$ and $d/n \to 0$ as $n \to \infty$. Then for any non-degenerate discrete memoryless channel C,

$$N_C(n, d, \varepsilon) \sim \frac{\ln \frac{n}{d+1} + \frac{1}{d+1} \ln \frac{1}{\varepsilon}}{\ln \frac{1}{\alpha(C)}}.$$

In the case when d grows linearly and ε decreases exponentially in n, a restricted (i.e., independent of n) number of repetitions are sufficient.

We also consider the optimization problems of repeated transmission over continuous channels with discrete time and additive noise [7]. The problem of finding the minimum number $N(n, \delta, \varepsilon)$ of repeated transmissions, which are sufficient for reconstruction of any real vector $x = (x_1, ..., x_n)$ within the Euclidean distance δ with error probability ε, is reduced to a classical statistical problem for the minimax estimate. In the case of a channel with the normal density, the optimal N-reconstructor is defined by $F(Y) = \frac{1}{N} \sum_{j=1}^{N} y_j$, and, using this N-reconstructor, N-tuple transmission of any x is equivalent to unitary transmission of x over the same channel with variance divided by N. We find $N(n, \delta, \varepsilon)$ for this channel and its asymptotic behavior.

Section V contains concluding remarks and open problems. Some results of the paper without proofs were announced in [15].

II. Combinatorial channels

We denote by F_q^n the set of sequences $x = (x_1, ..., x_n)$ over the alphabet $F_q = \{0, 1, \ldots, q-1\}$, $q \geq 2$. We shall also use for x the multiplicative writing $x = x_1 \cdots x_n$ considering x as a *word* of length n over the alphabet F_q. Let $F_q^* = \cup_{n=0}^{\infty} F_q^n$ be the set of all words over F_q. Every combinatorial channel will be characterized by a set H of one-to-one partial mappings $F_q^* \to F_q^*$. Elements of H are referred to as *single errors.* For any $t \in \{0, 1, ...\}$ and any $x \in F_q^*$ denote by $B_t(x, H)$ the set of all words $y \in F_q^*$ which can be obtained from x by at most t step-by-step applications of single errors from H. (This means that there exists an integer s, $0 \leq s \leq t$, words $z_1, ..., z_s, z_{s+1}$ where $z_1 = x$, $z_{s+1} = y$, and single errors $h_1, ..., h_s \in H$ such that h_i is defined at z_i and $h_i(z_i) = z_{i+1}$ for any $i = 1, ..., s$.) Given a set H of single errors, we understand a *combinatorial* (n,t)*-channel* as being a multiple-valued function which can map an arbitrary (input) word $x \in F_q^n$ to any (output) word from the set $B_t(x, H)$. If $V \subseteq F_q^n$ and $|B_t(x, H) \cap B_t(z, H)| = 0$ for any different $x, z \in V$, then V is an *error-correcting code* for the combinatorial (n,t)-channel and the problem to find its maximum size is one of the main problems in coding theory. However, in this paper we consider another problem. Let

$$N_H(V, t) = \max_{x,z \in V,\ x \neq z} |B_t(x, H) \cap B_t(z, H)|. \tag{2}$$

It is clear that for each $x \in V$, any $N = N_H(V,t) + 1$ different elements of the set $B_t(x, H)$ allow one to uniquely recover this x in the set V. On the other hand, (2) shows that $N = N_H(V,t) + 1$ is the minimum number having this property. It is worth pointing out that it is in general possible that $|B_t(x, H)| \leq N_H(V, t)$ for some $x \in V$. Thus, $N = N_H(V,t) + 1$ is the minimum number such that every $x \in V$ can be reconstructed with the help of any N *different* elements of $B_t(x, H)$ if they exist. This gives rise to the problem of finding $N_H(V, t)$ for different H, n, $V \subseteq F_q^n$, and t.

Now we give examples of the sets H of single errors and hence the corresponding combinatorial (n,t)-channels. We shall use a common name (for instance, substitutions or transpositions) for all elements of each such H which is usually referred to as the *type* of errors. The sets H are in general countable, and we denote by H_n the subset of all single errors of H which are defined on at least one element of F_q^n.

Example 2.1 (substitutions): $H = \{h_i^a : a \in F_q \backslash \{0\}, i = 1, 2, \ldots\}$ where the single error h_i^a is defined on all $x = x_1 \ldots x_n \in F_q^n$ such that $n \geq i$ and substitutes the letter x_i in x for the letter $x_i + a \bmod q$. In this case $|H_n| = (q-1)n$.

Example 2.2 (asymmetric errors): $H = \{h_i^> : i = 1, 2, \ldots\}$ and $H = \{h_i^< : i = 1, 2, \ldots\}$ where the single error $h_i^>$ $(h_i^<)$ is defined on all $x = x_1 \ldots x_n \in F_q^n$ such that $n \geq i$ and $x_i < q-1$ ($x_i > 0$) and substitutes the letter x_i in x for the letter $x_i + 1$ (respectively for $x_i - 1$). In the both cases $|H_n| = n$.

Example 2.3 (cyclic errors): $H = \{h_i^{\pm} : i = 1, 2, \ldots\}$ where the single error $h_i^{\pm}$ is defined on all $x = x_1 \ldots x_n \in F_q^n$, $q \geq 3$, such that $n \geq i$ and substitutes the letter x_i in x for the letter $x_i \pm 1 \bmod q$. In this case $|H_n| = 2n$.

Example 2.4 (transpositions): $H = \{h_{i,j} : i, j = 1, 2, \ldots, i < j\}$ where the single error $h_{i,j}$ is defined on all $x = x_1 \ldots x_n \in F_q^n$ such that $n \geq j$ and transposes the letters x_i and x_j in x. In this case $|H_n| = n(n-1)/2$.

Example 2.5 (deletions): $H = \{h_i^{a,-} : a \in F_q, i = 1, 2, \ldots\}$ where the single error $h_i^{a,-}$ is defined on all $x = x_1 \ldots x_n \in F_q^n$ such that $n \geq i$, $x_i = a$, and maps x to the word $x_1 \ldots x_{i-1} x_{i+1} \ldots x_n$ of length $n-1$. In this case $|H_n| = qn$.

Example 2.6 (insertions): $H = \{h_i^{a,+} : a \in F_q, i = 0, 1, \ldots\}$ where the single error $h_i^{a,+}$ is defined on all $x = x_1 \ldots x_n \in F_q^n$ such that $n \geq i$ and maps x to the word $x_1 \ldots x_i a x_{i+1} \ldots x_n$ of length $n+1$. In this case $|H_n| = q\,(n+1)$.

Other types of single errors such as bursts and arithmetic errors (see, for example, [17]) also admit a similar description.

An important problem is to find a simple algorithm for reconstruction of $x = x_1 \cdots x_n$ with the help of different words $y_1, ..., y_N \in B_t(x, H)$ where $N = N_H(V,t) + 1$. In the cases of single errors which change the length (for instance, deletions and insertions), we shall replace $B_t(x, H)$ in (2) by $B_t(x, H) \cap F_q^m$ for some m. Thus, we can assume that all words $y_1, ..., y_N$ have the same length m (usually $m = n$) and $y_j = y_{1,j} y_{2,j} \cdots y_{m,j}$, $j = 1, 2, ..., N$. Denote the matrix $(y_{i,j})$ of order $m \times N$ by Y and recall that all columns of Y are different. A problem is to give a simple description of the function (N-reconstructor) F such that $F(Y) = x$. To this end we introduce threshold functions defined on real vectors $w = (w_0, ..., w_{h-1}) \in R^h$ where $h = q$ or $h = t+1$. These vectors w are calculated by the matrix Y and do not depend on enumeration of its columns. We shall consider the alphabet $F_q = \{0, 1, ..., q-1\}$ as an ordered set. For any $u \in F_q^N$ denote by $k_i(u)$ the number of letters $i \in F_q$ in the word u and call

$$k(u) = (k_0(u), ..., k_{q-1}(u)) \tag{3}$$

the *composition* of u. Any word $u \in F_q^N$ *generates a permutation*

$$\theta(u) = (\theta_0, \theta_1, ..., \theta_{q-1}) \tag{4}$$

of F_q which is uniquely defined by the conditions: $k_{\theta_i}(u) \geq k_{\theta_j}(u)$ if $i < j$ and $\theta_i < \theta_j$ if $k_{\theta_i}(u) = k_{\theta_j}(u)$. We call

$$l(u) = (k_{\theta_0}(u), ..., k_{\theta_{q-1}}(u)) \tag{5}$$

the *ordered composition* of u. In particular, for $q = 3$ and

$$u = 0211012 \tag{6}$$

we have $k_0(u) = 2$, $k_1(u) = 3$, $k_2(u) = 2$, $k(u) = (2, 3, 2)$, $\theta(u) = (1, 0, 2)$, and $l(u) = (3, 2, 2)$. A function $m_q : F_q^N \to$

F_q is called the *majority function*, if $m_q(u) = \theta_0$ where θ_0 is defined by (4). If a letter $a \in F_q$ occurs in u more often than any other letter, then $m_q(u) = a$. For any real vector $\tau = (\tau_0, \tau_1, ..., \tau_{h-1}) \in R^h$ we consider the *threshold function* $f_{\tau_0,\tau_1,...,\tau_{h-1}} = f_\tau : R^h \to F_q$ (with thresholds $\tau_0, \tau_1, ..., \tau_{h-1}$) which is defined on all vectors $w = (w_0, w_1, ..., w_{h-1}) \in R^h$ with

$$\sum_{i=0}^{h-1} w_i > \sum_{i=0}^{h-1} \tau_i \tag{7}$$

and is equal to the minimum number $i \in F_q$ such that $w_i > \tau_i$. This definition is correct, since the condition (7) guarantees that such a number exists. Note that if $w_a > \tau_a$ is fulfilled only for a letter $a \in F_q$, then the threshold function is equal to a. Note also that the majority function $m_q(u)$ can be also expressed as $f_{0,...,0}(l(u))$ (with q zero thresholds). As an example, for (6) we have $m_3(u) = 1$, $f_{3,2,1}(k(u)) = 1$, and $f_{3,2,1}(l(u)) = 2$.

Considering a certain set H of single errors we shall omit H in notations $B_t(x, H)$. We assume that q, n, t are integers ($q \geq 2$, $n \geq 1$, $t \geq 0$) and put

$$\sum_{i=l}^{j} a_i = 0, \text{ if } j < l, \text{ and } \binom{t}{i} = 0, \text{ if } i < 0 \text{ or } i > t. \tag{8}$$

A. Substitutions

In the case when $V = F_q^n$ and H consists of qn single substitutions, the set $B_t(x)$ is the metric ball of radius t centered at point x of the Hamming space F_q^n (see Example 2.1). Put $N(F_q^n, t) = N_H(F_q^n, t)$ in this case. The problem is to find $N(F_q^n, t)$ and a simple algorithm of reconstruction of $x \in F_q^n$ with the help of any different $y_1, ..., y_N \in B_t(x)$ where $N = N(F_q^n, t) + 1$.

Theorem 1: For any n, q and t,

$$N(F_q^n, t) = q \sum_{i=0}^{t-1} \binom{n-1}{i} (q-1)^i. \tag{9}$$

Let different $y_1, ..., y_N$ belong to $B_t(x)$ for some $x = x_1 \cdots x_n \in F_q^n$, $N = N(F_q^n, t) + 1$, and $y_j = y_{1,j} \cdots y_{n,j}$, $j = 1, ..., N$. Then

$$x_i = m_q(y_{i,1}, ..., y_{i,N}),\ i = 1, ..., n. \tag{10}$$

Proof: Let all letters of words x, $z \in F_q^n$ coincide except the first one. Then all words which have an arbitrary first letter and differ from x (and z) in at most $t-1$ remaining places, belong to the set $B_t(x) \cap B_t(z)$. This proves that $N(F_q^n, t)$ is not smaller than the right-hand side of (9). The opposite inequality will follow from proof of the second part of Theorem. To this end we note that for any $i = 1, ..., n$ and $a \in F_q$, $a \neq x_i$, the number of $z = z_1 \cdots z_n \in B_t(x)$ such that $z_i = a$ equals $\sum_{j=0}^{t-1} \binom{n-1}{j} (q-1)^j$. Therefore, in the word $y_{i,1} \cdots y_{i,N} \in F_q^N$ the letter x_i occurs more often than others and we can apply the majority function to find x_i. □

Thus, Theorem 1 determines the minimum number $N = N(F_q^n, t)+1$ of different words obtained from a word $x \in F_q^n$ by at most t substitutions of symbols, which are sufficient for its reconstruction. Moreover, this indicates that such a reconstruction can be performed by applying the majority function to each row of the matrix $Y = (y_{i,j})$. Note that $N = q + 1$ and $N = q(q-1)(n-1) + q + 1$ for $t = 1$ and $t = 2$ respectively. A numeric example is given in the Introduction.

B. Transpositions and asymmetric errors

In the case when $V \subseteq F_q^n$ and H consists of $n(n-1)/2$ single transpositions, we put $N^\circ(V, t) = N_H(V, t)$ and note that any $z \in B_t(x)$ has the same composition as $x \in F_q^n$, i.e., $k(z) = k(x)$, see Example 2.4 and (3). In particular, for $V = F_2^n$ we can determine the Hamming weight w of an unknown binary word $x \in F_2^n$ by words of the set $B_t(x)$. Therefore it is natural to consider the problem to find $N^\circ(J_w^n, t)$ where J_w^n consists of all words $x \in F_2^n$ with w ones and $n - w$ zeros. If we consider the distance between elements of J_w^n being equal to a half of the Hamming distance (the Hamming distance is even in this case), then $B_t(x)$ is the metric ball of radius t centered at point $x \in J_w^n$. This metric space J_w^n is called the Johnson space.

Theorem 2: For any n, w and t,

$$N^\circ(J_w^n, t) = n \sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i} \frac{1}{i+1}. \tag{11}$$

Let different $y_1, ..., y_N$ belong to $B_t(x)$ for some $x = x_1 \cdots x_n \in J_w^n$, $N = N^\circ(J_w^n, t) + 1$, $y_j = y_{1,j} \cdots y_{n,j}$ (the j-th column), $j = 1, ..., N$, and $u_i = y_{i,1} \cdots y_{i,N}$ (the i-th row), $i = 1, ..., n$. Then

$$x_i = f_{\tau_0,\tau_1}(k(u_i)),\ i = 1, ..., n, \tag{12}$$

where

$$\tau_0 = \frac{n-w}{n} N^\circ(J_w^n, t) \text{ and } \tau_1 = \frac{w}{n} N^\circ(J_w^n, t).$$

Proof: Denote the right-hand side of (11) by N°. Let $z \in J_w^n$, $z \neq x$, be obtained from $x \in J_w^n$ by a single transposition $h_{l,m}$. Considering four possible cases for these two positions l and m it is easy to see that

$$|B_t(x) \cap B_t(z)| = 2 \sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i} +$$

$$\sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i+1} + \sum_{i=0}^{t-1} \binom{w-1}{i+1} \binom{n-w-1}{i} =$$

$$\sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w}{i+1} + \sum_{i=0}^{t-1} \binom{w}{i+1} \binom{n-w-1}{i} = N^\circ.$$

This proves that $N^\circ(J_w^n, t) \geq N^\circ$. The opposite inequality will follow from proof of the second part of Theorem. To prove this part we note that for any $x = x_1 \cdots x_n \in J_w^n$ and i, $i = 1, ..., n$, the number of $z = z_1 \cdots z_n \in B_t(x)$ such that $z_i \neq x_i$ equals

$$\sum_{i=0}^{t-1} \binom{w}{i+1} \binom{n-w-1}{i} = \frac{w}{n} N^\circ$$

if $x_i = 0$, and equals

$$\sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w}{i+1} = \frac{n-w}{n} N^\circ$$

if $x_i = 1$. This implies that for $N = N^\circ + 1$ the threshold function (12) allows us to find x_i and completes the proof of the fact that $N^\circ(J_w^n, t) = N^\circ$. □

Thus, Theorem 2 determines the minimum number $N = N^\circ(J_w^n, t) + 1$ of different words obtained from a word $x \in J_w^n$ by at most t transpositions of symbols which are sufficient for its reconstruction. Moreover, this indicates that such a reconstruction can be performed by applying a threshold function to the composition of each row of the matrix $Y = (y_{i,j})$. Note that $N = n + 1$ and $N = n(w(n-w) - n + 3)/2 + 1$ for $t = 1$ and $t = 2$ respectively. A numeric example is given in the Introduction.

It is worth pointing out that any $n+1$ words $y_j = y_{1,j} \cdots y_{n,j}$, $j = 1, ..., n+1$, obtained from an arbitrary $x = x_1 \cdots x_n \in F_q^n$ by at most one transposition are sufficient to reconstruct this x and hence $N^\circ(F_q^n, 1) = n$ for any $q \geq 2$. In this case we know the composition of the x (it coincides with that of all y_j) and can reconstruct x as follows:

$$x_i = f_{\tau_0, \tau_1, ..., \tau_{q-1}}(k(u_i)), \; i = 1, ..., n, \tag{13}$$

where $u_i = y_{i,1} \cdots y_{i,n+1}$ and

$$\tau_l = k_l(x), \; l = 0, ..., q-1.$$

It follows from the fact that the number of $z = z_1 \cdots z_n \in B_1(x)$ such that $z_i = a \in F_q^n$ and $a \neq x_i$ equals $k_a(x)$. Note that, in general, application of the majority function to all rows is not suitable in this case.

Example 2.7: Let the columns of the matrix

$$\begin{matrix} 0 & 0 & 1 & 1 & 1 & 2 & 2 & 2 \\ 2 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \\ 1 & 1 & 2 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 \\ 2 & 2 & 2 & 2 & 2 & 1 & 2 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 2 & 2 & 1 & 1 \end{matrix}$$

be obtained from an unknown $x \in F_3^7$ by at most one transposition of two symbols. Any column allows us to determine the following composition of the x : $k(x) = (2, 3, 2)$. Using (13) we find that $x = 2011201$.

Consider asymmetric errors $H = \{h_i^> : i = 1, 2, \dots\}$ and $H = \{h_i^< : i = 1, 2, \dots\}$ (see Example 2.2) for the set $V = J_w^n$ and denote $N_H(J_w^n, t)$ for these types of errors by $N^>(J_w^n, t)$ and $N^<(J_w^n, t)$ respectively.

Theorem 3: For any n, w and t,

$$N^<(J_w^n, t) = \sum_{i=0}^{t-1} \binom{w-1}{i}, \; N^>(J_w^n, t) = \sum_{i=0}^{t-1} \binom{n-w-1}{i}.$$

If different $y_1, ..., y_N$ belong to $B_t(x)$ for some $x = x_1 \cdots x_n \in J_w^n$, $N = N^>(J_w^n, t) + 1$ or $N = N^<(J_w^n, t) + 1$, and $y_j = y_{1,j} y_{2,j} \cdots y_{n,j}$, $j = 1, ..., N$, then

$$x_i = y_{i,1} \vee y_{i,2} \vee \cdots \vee y_{i,N}, \; i = 1, ..., n,$$

or respectively

$$x_i = y_{i,1} \& y_{i,2} \& \cdots \& y_{i,N}, \; i = 1, ..., n.$$

The proof of Theorem 3 is a simple version of that of Theorems 1 and 2. In these cases to reconstruct $x \in J_w^n$ one can apply conjunction or disjunction to each row of the matrix $Y = (y_{i,j})$.

C. Deletions and insertions

In the case when $V = F_q^n$ and H consists of qn single deletions and $q(n+1)$ single insertions (see Examples 2.5 and 2.6), the set $B_t(x)$ is the metric ball of radius t centered at $x \in F_q^n$ in the deletion/insertion metric introduced in [12] (see also [14]). Since the length of words of $B_t(x)$ varies from $n - t$ up to $n + t$, we consider separately the case of exactly t deletions and the case of exactly t insertions. Let

$$D_t(x) = B_t(x) \cap F_q^{n-t}, \qquad I_t(x) = B_t(x) \cap F_q^{n+t}. \tag{14}$$

It is obvious that $D_t(x)$ is the set of all words obtained from $x \in F_q^n$ by deletion of t letters (or its subsequences of length $n - t$) and $I_t(x)$ is the set of all words obtained from $x \in F_q^n$ by insertion of t letters (or its supersequences of length $n + t$). We find

$$N_q^-(n, t) = \max_{x, z \in F_q^n; \; x \neq z} |D_t(x) \cap D_t(z)| \tag{15}$$

and

$$N_q^+(n, t) = \max_{x, z \in F_q^n; \; x \neq z} |I_t(x) \cap I_t(z)|. \tag{16}$$

Moreover, we shall describe simple algorithms that recover $x \in F_q^n$ with the help of any $N_q^-(n, t) + 1$ different elements of $D_t(x)$ (if they exist) and with the help of any $N_q^+(n, t) + 1$ different elements of $I_t(x)$ (we verify that such a number of elements always exist in $I_t(x)$). The ground of these results is rather long and the detailed proofs will be published in Discrete Mathematics (in Russian; English translation: Discrete Mathematics and Applications).

For any nonnegative integers n and t define

$$D_q(n, t) = \max_{x \in F_q^n} |D_t(x)|.$$

It is useful to assume that $D_q(n, t) = 0$ for any integers n and t such that $n < t$ or $t < 0$ and to extend the definition of $D_q(n, t)$ to the case $q = 1$ when $F_q = \{0\}$. In this case $D_q(n, t) = 1$ if $n \geq t \geq 0$. For any permutation $\sigma = (\sigma_0, ..., \sigma_{q-1})$ of letters of F_q denote by $x_{\sigma,n}$ the word of length n which is a prefix of the word $(\sigma_0 \cdots \sigma_{q-1})^l$ where $l = \lceil n/q \rceil$. ($(\sigma_0 \cdots \sigma_{q-1})^l$ denotes concatenation of l words $\sigma_0 \cdots \sigma_{q-1}$.) It is clear that $|D_t(x_{\sigma,n})|$ does not depend on the permutation σ. As it was mentioned in [4] (with the reference to the report [3]) Calabi proved that

$$D_q(n, t) = |D_t(x_{\sigma,n})| \tag{17}$$

and found that for any nonnegative integer m,

$$\left(\sum_{j=0}^{q-1} z^j\right)^m \sum_{i=0}^{\infty} z^i = \sum_{h=0}^{\infty} D_q(m+h,h)z^h. \quad (18)$$

These results can be obtained with the help of the following two arguments:

$$D_q(n+1,t) \geq D_q(n,t-1), \quad (19)$$

and if $x_m = a$ is the first occurrence of a letter $a \in F_q$ in the word $x = x_1 \cdots x_n \in F_q^n$, then

$$|D_t^a(x)| \leq D_q(n-m, t-m+1) \quad (20)$$

where $D_t^a(x)$ is the set of all words of $D_t(x)$ which are starting from a. In particular, they show that for any $x \in F_q^n$ and any t, $n > t \geq 0$,

$$|D_t(x)| \leq \sum_{i=0}^{q-1} D_q(n-i-1, t-i)$$

with equality if $x = x_{\sigma,n}$ (the latter is proved by induction on n using the fact that any suffix of $x_{\sigma,n}$ is a prefix of $x_{\sigma',n}$ for some cyclic permutation σ' of σ). This gives (17) and (18), since for $D_q(t,t) = 1$ (18) is equivalent to the equality

$$D_q(n,t) = \sum_{i=0}^{q-1} D_q(n-i-1, t-i) \text{ when } n > t \geq 0. \quad (21)$$

Recently Hirschberg [10] reopened the Calabi results (17) and (21) and, for $q \geq 2$, $n \geq t \geq 0$, found the following important formula:

$$D_q(n,t) = \sum_{i=0}^{t} \binom{n-t}{i} D_{q-1}(t, t-i). \quad (22)$$

This gives the recurrence on q for computing $D_q(n,t)$ and, in particular, shows that

$$D_2(n,t) = \sum_{i=0}^{t} \binom{n-t}{i}, \; D_3(n,t) = \sum_{i=0}^{t} \binom{n-t}{i} \sum_{j=0}^{t-i} \binom{i}{j}. \quad (23)$$

(The formula for $q = 2$ is due to Calabi [3].) Note that (22) may be obtained using in (18) the expansion

$$\left(\sum_{j=0}^{q-1} z^j\right)^m = \sum_{i=0}^{m} \binom{m}{i} z^i \left(\sum_{j=0}^{q-2} z^j\right)^i.$$

One can show that for any n, t, and q, $n \geq t \geq 0$, $q \geq 2$,

$$N_q^-(n,t) = \sum_{i=1}^{q-1} D_q(n-i-1, t-i) + D_q(n-2, t-1). \quad (24)$$

Using (21) we have that for $n \geq t+1$

$$N_q^-(n,t) = D_q(n,t) - D_q(n-1,t) + D_q(n-2,t-1) \quad (25)$$

and for $n \geq t+2$

$$N_q^-(n,t) = \sum_{i=0}^{q-1} N_q^-(n-i-1, t-i). \quad (26)$$

From (22) and (25) it follows that

$$N_q^-(n,t) = \sum_{i=0}^{t-1} \binom{n-t-1}{t-i-1} \left(D_{q-1}(t,i) + D_{q-1}(t-1,i)\right), \quad (27)$$

and one can use (22) and (23) to calculate $N_q^-(n,t)$. In particular,

$$N_2^-(n,t) = 2\sum_{i=0}^{t-1} \binom{n-t-1}{i}, \quad (28)$$

$$N_3^-(n,t) = \sum_{i=0}^{t-1} \binom{n-t-1}{i} \sum_{j=0}^{t-i-1} \left(\binom{i+1}{j} + \binom{i}{j}\right) \quad (29)$$

$$N_q^-(n,0) = 0, \; N_q^-(n,1) = 2, \; N_q^-(n,2) = 2n-3-\delta_{q,2}, \quad (30)$$

$$N_q^-(n,3) = (n-2)^2 - (3n-10)\,\delta_{q,2} - \delta_{q,3} \quad (31)$$

where $\delta_{i,j}$ is the Kronecker symbol. Since $D_{q-1}(t,i) \leq (q-1)^{t-i}$, (27) also implies (1).

Now we describe an *algorithm for reconstruction* of an unknown $x = x_1 \cdots x_n \in F_q^n$ when we know a set $\{y_1, ..., y_N\} \subseteq D_t(x)$ (i. e., N of its different subsequences of length $n-t$) where $N = N_q^-(n,t) + 1$ and $n > t \geq 1$. Note that in this case in fact $n > t+1$, since

$$N_q^-(t+1,t) = \min(t+1,q) \text{ for } t \geq 1$$

and $N_q^-(t+1,t) \geq |D_t(x)|$ for any $x \in F_q^{t+1}$. The algorithm consists in successive application of threshold functions to the ordered composition of the first rows of the matrices formed by erroneous patterns. At any step the first letter and also j, $0 \leq j \leq t$, deleted letters of the unknown word x are determined, and the problem is reduced to a similar analysis of a submatrix with the smaller number of rows.

We shall consider the words $y_j = y_{1,j}y_{2,j} \cdots y_{n-t,j}$, $j = 1, ..., N$, as columns of a matrix Y of order $(n-t) \times N$. For any $a \in F_q$, denote by Y_a the submatrix of Y formed by all of its columns whose first letter is a and by Y_a' the submatrix of Y_a which is obtained by removing the first row of Y_a. For the first row $u = y_{1,1}y_{1,2} \cdots y_{1,N}$ of Y find the permutation $\theta(u) = (\theta_0, \theta_1, ..., \theta_{q-1})$ and the ordered composition $l(u) = (k_{\theta_0}(u), ..., k_{\theta_{q-1}}(u))$ of u (see (4) and (5)). Consider the thresholds

$$\tau_i = N_q^-(n-i-1, t-i), \; i = 0, 1, ..., q-1, \quad (32)$$

and note that due to (26) the threshold function $f_{\tau_0,\tau_1,...,\tau_{q-1}}$ is defined on $l(u)$. If $f_{\tau_0,\tau_1,...,\tau_{q-1}}(l(u)) = \theta_j$, then $0 \leq j \leq t$ and

$$x_i = \theta_{i-1} \text{ for } i = 1, ..., j+1. \quad (33)$$

When $j = t$ the submatrix Y'_{θ_j} consists of the unique column $x_{j+2}\cdots x_n$, and reconstruction of x is completed. When $j < t$ all columns of Y'_{θ_j} belong to $D_{t-j}(x_{j+2}\cdots x_n)$ and their number exceeds $N_q^-(n-1-j, t-j)$, and hence the problem is reduced to that of reconstructing the word $x' = x_{j+2}\cdots x_n$ of the smaller length $n-j-1$ by $N_q^-(n-j-1, t-j)+1$ its different subsequences of length $n-t-1$ (obtained from x' by $t-j \geq 1$ deletions).

Thus, the first letter of x is recovered with the help of application of the majority function to the first row u of the matrix Y, since $x_1 = \theta_0 = m_q(u)$. We then go to analysis of the largest submatrix Y'_{θ_0} only if Y'_{θ_0} is too large or precisely if the number of its columns exceeds the threshold $N_q^-(n-1, t)$. Otherwise, we need to investigate the smaller submatrix Y'_{θ_j} where j is defined as above so as not to lose the required information on x.

Example 2.8 : Let $q = 3$, $n = 7$, $t = 3$, $N_3^-(7,3) = 24$, by (31), and 25 columns of the matrix

$$\begin{matrix}
2&2&2&2&2&2&2&2&2&2&2&2&2&2&0&0&0&0&0&0&0&0&1&1&1&1\\
0&0&0&0&0&0&0&0&1&1&1&1&2&1&1&1&1&1&1&2&2&0&2&2&2&0\\
1&1&1&2&2&2&0&0&2&2&0&0&0&2&2&2&0&0&0&0&2&0&0&2&2\\
0&1&2&0&1&2&1&2&0&1&1&2&1&0&1&2&1&2&1&2&1&1&2&1&1
\end{matrix}$$

be subsequences of a certain $x = x_1\cdots x_7 \in F_3^7$. For the first row $u \in F_3^{25}$ of the matrix we have $k_0(u) = 8$, $k_1(u) = 4$, $k_2(u) = 13$, and hence $\theta(u) = (2,0,1)$ and $l(u) = (13,8,4)$. Since $\tau_0 = N_3^-(6,3) = 15$, $\tau_1 = N_3^-(5,2) = 7$, $\tau_2 = N_3^-(4,1) = 2$, we have $f_{\tau_0,\tau_1,\tau_2}(l(u)) = \theta_1 = 0$. Hence $x_1 = 2$, $x_2 = 0$ and the problem is reduced to reconstruction from $x_3\cdots x_7 \in F_3^5$ by $N_3^-(5,2)+1 = 8$ of its subsequences of length 3 which are columns of the following matrix Y'_0 :

$$\begin{matrix}
1&1&1&1&1&2&2&0\\
2&2&2&0&0&0&0&2\\
0&1&2&1&2&1&2&1
\end{matrix}$$

Now for the first row $u \in F_3^8$ of this matrix we have $k_0(u) = 1$, $k_1(u) = 5$, $k_2(u) = 2$, and hence $\theta(u) = (1,2,0)$ and $l(u) = (5,2,1)$. Since $\tau_0 = N_3^-(4,2) = 5$, $\tau_1 = N_3^-(3,1) = 2$, $\tau_2 = N_3^-(2,0) = 0$, we have $f_{\tau_0,\tau_1,\tau_2}(l(u)) = \theta_2 = 0$. Hence $x_3 = 1$, $x_4 = 2$, $x_5 = 0$. In this case $j = t = 2$ and the last column determines the remaining letters $x_6 = 2$, $x_7 = 1$. Thus, we finally have $x = 2012021$.

Considering insertions we first should mention the surprising fact that $|I_t(x)|$ does not depend on $x \in F_q^n$ and

$$|I_t(x)| = I_q(n,t) \quad \text{for any } x \in F_q^n \tag{34}$$

where

$$I_q(n,t) = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i. \tag{35}$$

This result was published in [13] for $q = 2$; the extension to the general case is immediate. We shall also assume that $I_q(n,t) = 0$ for $n \geq 0$ and $t < 0$, and that agrees with (35) (see (8)). Since for any $x = x_1\cdots x_n \in F_q^n$,

$$I_t(x) = x_1 I_t(x_2\cdots x_n) \bigcup \bigcup_{a\in F_q\setminus\{x_1\}} aI_{t-1}(x_1\cdots x_n) \tag{36}$$

(here $aU = \{z = au : u \in U\}$ for any $a \in F_q$ and any $U \subseteq F_q^{n+t-1}$), this fact can be proved by induction on $n+t$ ($n \geq 1, t \geq 1$), and (35) can be easily found by calculation of $|I_t(0^n)|$. Note that from (36) it is also follows that for $n \geq 1$

$$I_q(n,t) = I_q(n-1,t) + (q-1)I_q(n,t-1) \tag{37}$$

and hence

$$I_q(n,t) = \sum_{i=0}^{t} I_q(n-1, t-i)(q-1)^i \tag{38}$$

because $I_q(n,0) = 1$ for any $n \geq 0$.

One can show that for any n, t, and q, $n \geq t \geq 0$, $q \geq 2$,

$$N_q^+(n,t) = qI_q(n,t-1) - Q_q(n,t) \tag{39}$$

where

$$Q_q(n,t) = \sum_{i=0}^{t-1} \binom{n-1+i}{n-1} (q-1)^i (q-2)^{t-i}. \tag{40}$$

($Q_q(n,t)$ and $N_q^+(n,t)$ equal 0 if $t \leq 0$.) It is interesting that from (40) it follows that for $n \geq 1$

$$Q_q(n,t) = Q_q(n-1,t) + (q-1)Q_q(n,t-1) \tag{41}$$

and hence

$$Q_q(n,t) = \sum_{i=0}^{t} Q_q(n-1,t-i)(q-1)^i \tag{42}$$

since $Q_q(n,0) = 0$ for any $n \geq 0$. Thus, $I_q(n,t)$ and $Q_q(n,t)$ differ in only the choice of the initial conditions in (37) and (41): $I_q(n,0) = 1$, $I_q(0,t) = q^t$, $Q_q(n,1) = q-2$, $Q_q(1,t) = (q-2)\ ((q-1)^t - (q-2)^t)$. Combining (38), (39), and (42) we get

$$N_q^+(n,t) = \sum_{i=0}^{t} N_q^+(n-1,t-i)(q-1)^i \text{ for } n \geq 1. \tag{43}$$

Note that for $n \geq 1$, $t \geq 0$,

$$I_q(n,t) - qI_q(n,t-1) = I_q(n-1,t) - I_q(n,t-1)$$
$$= \binom{n-1+t}{t} (q-1)^t > 0,$$

by (35) and (37), and hence for any $x \in F_q^n$, $|I_t(x)| > N_q^+(n,t)$.

Now we describe an *algorithm for reconstruction* of an unknown $x = x_1\cdots x_n \in F_q^n$ with the help of any set $\{y_1, ..., y_N\}$ from $N = N_q^+(n,t)+1$ different elements of $I_t(x)$ (i. e., its supersequences of length $n+t$), $n \geq 1$, $t \geq 1$. We consider the words $y_j = y_{1,j}y_{2,j}\cdots y_{n+t,j}$, $j = 1, ..., N$, as columns of a matrix Y of order $(n+t)\times N$. For any $a \in F_q$ and $i = 1, 2, ..., n+t$, denote by $Y_{a,i}$ the submatrix of Y consisting of all columns $y_j = y_{1,j}y_{2,j}\cdots y_{n+t,j}$ such that $y_{i,j} = a$ is the first occurrence of the letter a in the word y_j, and denote by $Y'_{a,i}$ a submatrix of $Y_{a,i}$ which consists of all different columns obtained from $Y_{a,i}$ by removing the first i rows. Note that if $|U|$ denotes the number of columns of a matrix U, then $\left|Y'_{a,i}\right| \geq |Y_{a,i}|\,(q-1)^{-i+1}$. For any $b \in F_q$ consider the vector $w(b) \in R^{t+1}$ defined as follows:

$$w(b) = (w_0(b), ..., w_t(b)) \quad \text{where } w_i(b) = |Y_{b,i+1}|\,.$$

It is clear that

$$\sum_{i=0}^{t} w_i(x_1) = N_q^+(n,t) + 1. \tag{44}$$

On the other hand, one can show that for any $b \in F_q$, $b \neq x_1$,

$$\sum_{i=0}^{t} w_i(b) \leq N_q^+(n,t).$$

This allows us to find x_1 as the unique letter $b \in F_q$ for which $\sum_{i=0}^{t} w_i(b) > N_q^+(n,t)$. Consider the thresholds

$$\tau_i = N_q^+(n-1, t-i)(q-1)^i, \; i = 0, 1, ..., t,$$

and note that due to (43) and (44) the threshold function $f_{\tau_0,\tau_1,...,\tau_t}$ is defined on $w(x_1) \in R^{t+1}$. If $f_{\tau_0,\tau_1,...,\tau_t}(w(x_1)) = j$, then $0 \leq j \leq t$ and the set $Y_{x_1,j+1}$ contains at least $N_q^+(n-1,t-j)(q-1)^j$ +1 different columns and hence $Y'_{x_1,j+1}$ contains at least $N_q^+(n-1, t-j)+1$ different columns each of which belongs to $I_{t-j}(x_2\cdots x_n)$. When $j = t$, $Y'_{x_1,j+1}$ consists of the only column $x_2\cdots x_n$ and this completes the reconstruction of the x. When $0 \leq j < t$ we determine x_1 and reduce the problem considered to reconstruction of $x' = x_2\cdots x_n$ from any $N_q^+(n-1,t-j)+1$ different words of the set $I_{t-j}(x_2\cdots x_n)$.

Thus, the first letter x_1 of x is recovered with the help of calculation of vectors $w(b) \in R^{t+1}$, $b \in F_q$. In general, this letter differs from the letter obtained by applying the majority function to the first row of Y. The number $j = f_{\tau_0,\tau_1,...,\tau_t}(w(x_1))$, $0 \leq j \leq t$, is equal to the number of inserted letters. In order to find x_2 we need to investigate $Y'_{x_1,j+1}$ and so on. Note also that (39), (40) and (35) imply that $N_q^+(n,0) = 0$, $N_q^+(n,1) = 2$, $N_q^+(n,2) = 2(q-1)(n+2)$.

Example 2.9: Let $q = 3$, $n = 3$, $t = 2$, and hence $N_3^+(3,2) = 20$. Consider the following matrix Y derived from 21 supersequences of an unknown $x = x_1x_2x_3 \in F_3^3$:

$$\begin{matrix}
0&0&0&0&0&0&0&0&0&1&1&1&1&1&1&1&1&1&2&2&2\\
0&1&2&2&2&2&2&2&2&0&1&2&2&2&2&2&2&2&0&1&2\\
2&2&0&0&0&0&0&1&2&2&2&0&0&0&0&0&1&2&2&2&2\\
0&0&1&1&0&2&1&0&0&0&0&1&1&0&2&1&0&0&0&0&0\\
1&1&1&2&1&1&0&1&1&1&1&1&2&1&1&0&1&1&1&1&1
\end{matrix}$$

We see that $w(0) = (9,2,5)$, $w(1) = (9,2,1)$, $w(2) = (3,14,4)$, and hence $x_1 = 2$. Since $\tau_0 = 16$, $\tau_1 = 4$, $\tau_2 = 0$, we have $f_{16,4,0}(w(2)) = 1$. The matrix $Y_{2,2}$ has 14 columns and $Y'_{2,2}$ has 7 different columns each of which is obtained from x_2x_3 by a single insertion. However, for the further reconstruction it is sufficient to use any $N_3^+(2,1)+1 = 3$ of them, for instance,

$$\begin{matrix}
0&0&2\\
1&2&0\\
2&1&1
\end{matrix}$$

We find $w(0) = (2,1)$, $w(1) = (0,1)$, $w(2) = (1,1)$, and hence $x_2 = 0$. Now $\tau_0 = 2$, $\tau_1 = 0$, $f_{2,0}(w(0)) = 1$, and $Y'_{0,2}$ consists from one letter $x_3 = 1$. Thus, $x = 201$.

III. Graph-theoretical approach to reconstruction of sequences

The considered combinatorial problems of efficient reconstruction of sequences can be reduced to some extremum problems of reconstruction of vertices of graphs using the minimum number of different vertices in their metrical balls of a restricted radius. The problems of exact reconstruction of a vertex and its reconstruction with a preset accuracy expressed in terms of the path distance of a graph are given. We introduce the property of a graph to be monotone on intersections. This property of a graph allows us to find solutions of the problems in terms of parameters of the graph. The problem of reconstruction of vertices within distance d is equivalent to exact reconstruction of vertices belonging to a subset of vertices (code) with the minimum distance d between its different elements. Together with the known problems of exact reconstruction of vertices of codes with the given minimum distance this gives rise to new problems when the minimum distance of a code is less than the required accuracy of reconstruction. Although these problems have the sense for an arbitrary graph, we consider a special representation of a graph based on the description of its path metric with the help of a set H of single errors which are partial one-to-one mappings on the set of vertices. We show that this approach does not lose generality. On the other hand, many types of errors of essential interest in coding theory imply the natural description of the corresponding graphs. Moreover, this approach allows us to formulate a sufficient condition for a graph to be monotone on intersections. We use this condition in order to find solutions of the extremum problems for some types of single errors.

A. Reconstructing vertices of a graph

Let $\Gamma = \{V, E\}$ be a *directed graph* with a finite set V of vertices and a set $E \subseteq V \times V$ of edges. Thus, we do not consider the parallel edges. Moreover, we assume that E does not contain loops (x,x), $x \in V$. When every two ordered pairs (x,y) and (y,x) belong or do not belong to E simultaneously, we simply call $\Gamma = \{V, E\}$ a *graph* and identify every $(x,y) \in E$ with $(y,x) \in E$. We denote by $\rho(x,y)$ the *path metric* of a graph Γ being equal to the minimum number of edges in a path joining x and y. We do not in general assume that Γ is a connected graph and put $\rho(x,y) = \infty$ if x and y belong to different components. We denote by $s = s(\Gamma)$ the maximum of $\rho(x,y)$ over all $x, y \in V$ such that there exists a path joining x and y. This is called *diameter* of Γ in the case of connected Γ. For any

$x \in V$ and i, $i = 0, 1, ..., s$, we consider the *metrical spheres of radius* i

$$S_i(x) = \{y : y \in V,\ \rho(x, y) = i\} \tag{45}$$

(centered at x) and the *metrical balls of radius* i

$$B_i(x) = \bigcup_{j=0}^{i} S_j(x).$$

The *degree* $r = r(\Gamma)$ of a graph Γ is the maximum of $|S_1(x)|$ over all $x \in V$. A graph Γ is called *regular* of degree r if $|S_1(x)| = r$ for all $x \in V$. For any $i, j \in \{0, 1, \ldots, s\}$ and $x, y \in V$, let

$$p_{i,j}(x, y) = |S_i(x) \bigcap S_j(y)|. \tag{46}$$

Connected graphs Γ for which values (46) depend only on i, j and $\rho(x, y)$ are called *distance-regular*. For such graphs, we denote $p_{i,j}(x, y)$ by $p^k_{i,j}$ if $\rho(x, y) = k$. There exists one-to-one correspondence between distance-regular graphs of diameter s and symmetric association schemes with s classes [5]. An *automorphism* of a graph Γ is a permutation g of the vertex set V such that $(x, y) \in E$ if and only if $(g(x), g(y)) \in E$. For any $x, y \in V$ and an automorphism g, $\rho(x, y) = \rho(g(x), g(y))$. A graph Γ is called *distance-transitive* if for any $x, y, x', y' \in V$ such that $\rho(x, y) = \rho(x', y')$ there exists an automorphism g of Γ for which $x' = g(x)$, $y' = g(y)$. Distance-transitive graphs are distance-regular.

For a graph Γ we define some functions of essential interest for the problems of reconstructing vertices. For any t, $k, d \in \{0, 1, ..., s\}$ put

$$m_k(V; t) = \max_{x,y \in V,\ \rho(x,y)=k} |B_t(x) \cap B_t(y)| \tag{47}$$

and consider the main value

$$N(V; t, d) = \max_{d \le k \le s} m_k(V; t). \tag{48}$$

This value is equal to the maximum number of vertices in the intersection of two metric balls of radius t with centers at distance d or more one from another. In particular, $N(V; t, d) = 0$ for $d \ge 2t + 1$. The important point to note here is the significance of $N(V; t, d)$ for the following problem of approximation of an arbitrary vertex x of a subset (code) $C \subseteq V$ by some $z \in V$ such that $\rho(x, z)$ does not exceed a preset distance ρ using the minimum number of vertices of $B_t(x)$.

Fix integers t and ρ, $0 \le \rho \le t$, and a code $C \subseteq V$. To avoid trivial cases, we assume that $B_t(x)$ and $B_t(y)$ are different for different $x, y \in V$ and C does not belong to $B_\rho(z)$ with $z \in V$. Let $M(C; t, \rho)$ denote the minimum integer M, such that for any set $Y \subseteq V$, for which $|Y| = M$ and the set

$$A_t(C, Y) = \{x : x \in C,\ Y \subseteq B_t(x)\} \tag{49}$$

(of possible centers) is not empty, there exists a vertex $z \in V$ (permissible approximation) such that $\rho(x, z) \le \rho$ for all $x \in A_t(C, Y)$. This means (in particular, for $C = V$) that $M(C; t, \rho)$ is the minimum size of a set Y of vertices (erroneous patterns) in metric ball $B_t(x)$ centered at an arbitrary $x \in C$ which allows one to approximate this x by an vertex $z \in V$ to within distance ρ (or exactly when $\rho = 0$). Under our assumptions $1 \le M(C; t, \rho) \le \max_{x \in V} |B_t(x)|$. Note that $M(C; t, t) = 1$ for any $C \subseteq V$, since in case $\rho = t$ any element of $Y \subseteq B_t(x)$ can be chosen as permissible approximation.

Lemma 1: For any integers t and ρ, $0 \le \rho \le t$, $2\rho + 1 \le s$,

$$M(V; t, \rho) \ge N(V; t, 2\rho + 1) + 1 \tag{50}$$

with equality when $\rho = 0$ or $\rho = t$.

Proof: Case $\rho = t$ has been considered. For $0 \le \rho < t$ there exist $x, y \in V$ such that $\rho(x, y) \ge 2\rho + 1$ and $x, y \in A_t(V, Y)$ where $Y = B_t(x) \cap B_t(y)$ and $|Y| = N(V; t, 2\rho + 1) \ge 1$. This contradicts to the existence $z \in V$ for which $\rho(x, z) \le \rho$ for all $x \in A_t(V, Y)$ and gives (50). On the other hand, if $|Y| \ge N(V; t, 2\rho + 1) + 1$ for some $Y \subseteq V$, then the distances between elements of $A_t(V, Y)$ (if they exist) are not greater than 2ρ. Therefore, for $\rho = 0$ the set $A_t(V, Y)$ contains at most one element and we have equality in (50). □

Thus,

$$M(V; t, 0) = N(V; t, 1) + 1 \tag{51}$$

is equal to the minimum number of vertices in the metric ball of radius t with the center at an arbitrary $x \in V$ that are sufficient to exactly reconstruct this vertex. Note that, in general, for a set $A \subseteq V$ with pairwise distances $\le 2\rho$ one cannot guarantee the existence of a vertex $z \in V$ such that $\rho(x, z) \le \rho$ for all $x \in A$. Nevertheless, one may expect that for some graphs, in particular, for the Hamming and Johnson graphs, equality in (50) takes place for all ρ.

The value $N(V; t, d) + 1$ has one more natural interpretation connected with exact reconstruction of vertices of a code $C \subseteq V$ with the minimum distance

$$d(C) = \min_{x,y \in C, x \ne y} \rho(x, y). \tag{52}$$

Lemma 2: For any t, $d \in \{0, 1, ..., s\}$ and any code $C \subseteq V$ such that $d(C) \ge d$,

$$M(C; t, 0) \le N(V; t, d) + 1 \tag{53}$$

with equality for some codes.

Proof: In trivial case $d \ge 2t + 1$ we have $N(V; t, d) = 0$, $M(C; t, 0) = 1$ and any vertex of $B_t(x)$ allows one to exactly reconstruct $x \in C$. By the definition of $N(V; t, d)$, for any $Y \subseteq V$ such that $|Y| = N(V; t, d) + 1$, the set (49) can contain only one vertex of a code $C \subseteq V$ with $d(C) \ge d$. This implies (53). On the other hand, in the case $d \le 2t$ there exist $x, y \in V$ such that $\rho(x, y) \ge d$ and

$$1 \le |B_t(x) \bigcap B_t(y)| = N(V; t, d)$$

Therefore, if these x and y belong to a code C, then they also belong to (49) where $Y = B_t(x) \bigcap B_t(y)$ and hence $M(C; t, 0) > N(V; t, d)$. □

Lemmas 1 and 2 shed light on the role of $N(V;t,d)$, in particular, for $d=1$, for some new problems of coding theory when the both correcting ability of codes and repeated transmission of messages are used. For calculation of $N(V;t,d)$ we use a property of graphs which reflects our intuitive expectation that $m_k(V;t)$ must decrease with increasing k. A graph Γ is called *monotone on intersections* if for any $t, t=1,...,s$, value $m_k(V;t)$ does not increase with $k, k=1,...,s$. If Γ has this property, then

$$N(V;t,d)=m_d(V;t)=\max_{x,y\in V,\rho(x,y)=d}\sum_{i=0}^{t}\sum_{j=0}^{t}p_{i,j}(x,y).$$

However, we shall see that there exist "very symmetric" (in particular, distance-transitive) graphs which do not have this property.

Using arguments of the proof of Lemma 2 we get the following statement.

Lemma 3: If a distance-regular graph $\Gamma=\{V,E\}$ is monotone on intersections, then

$$N(V;t,d)=\sum_{i=0}^{t}\sum_{j=0}^{t}p_{i,j}^{d}$$

and for any code $C\subseteq V$ such that $d(C)=d$,

$$M(C;t,0)=N(V;t,d)+1.$$

B. Graphs with error metric

The metric approach to the problem of efficient reconstruction of sequences for different types of errors gives rise to the natural definition of a class of graphs including the Cayley graphs. Let V be a finite (or countable) set. Consider a set H of one-to-one, in general, partial mappings $V\to V$ which are referred to as *single errors*. We assume that V and H have the following property: if $h\in H$ is defined on $x\in V$ and $h(x)=y\in V$, then there exists $g\in H$ which is defined on y and $g(y)=x$. We shall write $H=H^{-1}$ if this property is satisfied. If V is finite and $h\in H$ is defined on all V, then the single error h is simply a *permutation* of V. A single error $h\in H$ being a permutation is called an *involution* if $h^{-1}=h$. It is clear that $H=H^{-1}$ if H together with any permutation h of V contains h^{-1}, in particular, if H consists of involutions of V. Let us construct a graph $\Gamma_H=\Gamma(V,E)$ (without loops and parallel edges) with the set V of vertices and the set E of edges, where $(x,y)\in E$ if and only if $x\neq y$ and there exists $h\in H$ such that $y=hx$. Here and subsequently, hx stands for the $h(x)$. This construction and property $H=H^{-1}$ imply that the path metric $\rho(x,y)$ on graph Γ_H is equal to the minimum number of single errors transforming x to y, if there exists a chain of such mappings, or ∞, otherwise. We call Γ_H a *graph with error metric* (of the type H).

Note that one can assume without loss of generality that the set H of single errors does not contain the identity and for any $h\in H$ and $x\in V$ either h is not defined on x or $hx\neq x$. This follows from the fact that the definition of Γ_H does not depend on whether $hx=x$ or h is not defined on x.

The construction of the graphs $\Gamma_H=\Gamma(V,E)$ with $V=F_q^n$ can be used for many types $H=H_n$ of single errors considered in coding theory (see also [12], [13], [17]). In case $H=H^{-1}$ the set $B_i(x,H)$ defined in Section II coincides with the metric ball $B_i(x)$ in Γ_H. In particular, for the set $H=H_n$ of substitutions (Example 2.1) and cyclic errors (Example 2.3) we have $H=H^{-1}$ and get graphs with the Hamming and Lee metric respectively; for $V=J_w^n$ and the set $H=H_n$ of transpositions (Example 2.4) being involutions we also have $H=H^{-1}$ and get the graph with the Johnson metric. In Section II-A and II-B we in fact calculated $N(V;t,1)=N_H(V,t)$ for these cases. For the union H (countable set) of deletions and insertions on F_q^* (Examples 2.5 and 2.6) the required property $H=H^{-1}$ is satisfied, and the graph Γ_H generates on F_q^* the deletion/insertion metric [12]. Any $x\in F_q^*$ belongs to the range of definition of a finite number of $h\in H$ and hence all metric spheres (45) are finite for this countable graph. In this case we found

$$N_q^{\mp}(n,t)=\max_{x,z\in F_q^n;\ x\neq z}\left|B_t(x)\cap B_t(z)\cap F_q^{n\mp t}\right|.$$

The restriction $H=H^{-1}$ in the definition of $\Gamma_H=\Gamma(V,E)$ can be weakened. Let the following *parallelogram property* hold: for any $x,y,z\in V$ and $h,g\in H$ such that $x=hz$, $y=gz$ there exist $z'\in V$ and $h',g'\in H$ for which $z'=h'x$, $z'=g'y$. If for any $x,y\in V$ we put $(x,y)\in E$ if and only if $x\in B_1(y,H)$ or $y\in B_1(x,H)$, then one can show that the path metric of the graph $\Gamma_H=\Gamma(V,E)$ equals $\min(i+j)$ where the minimum is taken over all i and j such that there exists $z\in V$ for which $z\in B_i(x,H)$ and $z\in B_j(y,H)$. In particular, for the sets H of asymmetric errors on F_q^* (Example 2.2) it is not true that $H=H^{-1}$. However, in the both cases the parallelogram property holds, and we get the same graph Γ_H on the set F_q^*. In Section II-B for this graph we in fact found

$$\max_{x,z\in J_w^n;\ x\neq z}|B_t(x)\cap B_t(z)\cap J|$$

where $J=\bigcup_{i=0}^{w}J_i^n$ and $J=\bigcup_{i=w}^{n}J_i^n$ respectively.

An important class of graphs $\Gamma_H=\Gamma(V,E)$ with error metric is obtained when V is a finite group, a subset $H^\circ\subseteq V$ contains h^{-1} if $h\in H^\circ$, and H consists of all left (for definiteness) multiplications of elements H° by elements of V. We shall not distinguish elements of H and H° and shall write $H=H^\circ$. In this case all single errors of H are permutations of V. They all are automorphisms of Γ_H when V is an Abelian group. A graph $\Gamma_H=\Gamma(V,E)$ with Abelian group V and $H\subseteq V$ is referred to as *Abelian.* Many types of errors, such as substitutions, bursts, cyclic and arithmetic errors, give rise to Abelian graphs Γ_H.

Lemma 4: Any finite graph Γ of degree r with v vertices is a graph Γ_H with error metric for which H consists of involutions and

$$|H|\leq 2\left\lfloor\frac{v+1}{2}\right\rfloor-1 \text{ or } |H|\leq r+1\ . \qquad (54)$$

Each of these bounds is attained for any $v \geq 2$ and $r \geq 2$.

Proof: A graph $\Gamma = \Gamma(V, E)$ (without loops and parallel edges) coincides with a graph Γ_H where H consists from m involutions if and only if there exists a partition of Γ into m subgraphs whose vertices have degree at most one. Thus, this is equivalent to the known problem of coloring edges of a graph using the minimum number m of colors for the colors of any adjacent edges to be different. In the case of the complete graph Γ^v with v vertices it is known (see for example [9]) that this minimum number equals v, if v is odd, and $v-1$, if v is even. (Some readers felt this fact on practice participating in chess tournaments!) Since any graph can be extended to the complete graph with the same number of vertices, this gives the first bound in (54) and the fact that this can be tight for any v. The problem of coloring a graph of degree r with parallel edges was solved by Shannon in [20]. He proved that $m \leq \lfloor 3r/2 \rfloor$ and this bound is tight for any $r \geq 2$. This problem for the considered class of graphs without parallel edges was solved by Vizing in [22]. He proved that $m \leq r+1$ and the fact that this bound is attained for any $r \geq 2$. □

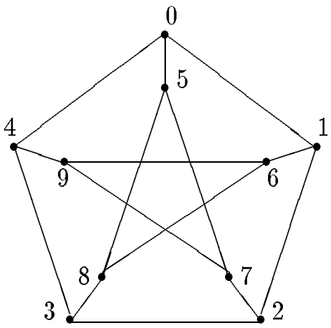

Fig. 2. The Peterson graph

For odd v and any r the bounds (54) are improved up to the minimum values $v-1$ and r if H can contain single errors which are not necessarily involutions of V. To prove the fact, we treat Γ as a directed graph replacing an edge jointing $x \in V$ with $y \in V$ by the pair of directed edges (x, y) and (y, x). Note that any directed subgraph of Γ, whose nonisolated vertices have the input and output degree at most one (or exactly one and, hence, belong to a directed cycle), defines a partial one-to-one mapping on V (respectively, a permutation of V). Therefore, a graph Γ coincides with a graph Γ_H where H consists from m single errors (permutations) if and only if there exists a partition of Γ into m directed subgraphs whose nonisolated vertices have the input and output degree at most one (exactly one, in the case of permutations). We shall consider regular graphs Γ, since any graph can be converted to a regular graph of the same degree adding vertices and edges. For odd v the complete graph Γ^v is partitioned into $(v-1)/2$ Hamilton cycles and hence into $v-1$ Hamilton directed cycles. This gives $m = v-1$ single errors (permutations, in the case of complete graphs) for odd v as well. By the Peterson theorem (see for example [9]), any regular graph of even degree r is partitioned into $r/2$ subgraphs of degree two. This gives $m = r$ single errors (permutations, in the case of regular graphs) in the case of even r. The following elegant proof valid for even and odd r was proposed by A. Brouwer for discussion of the problem. The *bipartite double* [2] of $\Gamma = \Gamma(V, E)$ is the graph Γ' with the vertex set $V' = \{\gamma^a : \gamma \in V,\ a \in \{0, 1\}\}$, such that its vertices γ^a and ξ^b are adjacent if and only if γ and ξ are adjacent in Γ and $a \neq b$). If Γ is a regular graph of degree r, then Γ' is a bipartite regular graph of degree r and, hence, there exists a partition of Γ' into r regular subgraphs of degree one (1-factors). Replace any edge jointing $\gamma^a \in V'$ with $\xi^b \in V'$ by the directed edge (γ^a, ξ^b) when $a = 0$ or by the directed edge (ξ^b, γ^a) when $b = 0$ and identify the vertices γ^0 and γ^1 of Γ'. As a result we obtain the directed version of the graph Γ. Moreover, the partition of the graph Γ' into r 1-factors generates a partition of the directed graph Γ into r directed subgraphs each of which consists of disjoint cycles since every of its vertices has the input and output degree one. Thus, the following statement holds.

Lemma 5: Any finite graph Γ of degree r is a graph Γ_H with error metric for which H consists of r single errors being permutations of vertices in the case of regular Γ.

The graphs in Fig. 2 and Fig. 3 are examples of regular graphs of degree 3 which cannot be represented as graphs Γ_H with three involutions. However, by Lemma 5, there exist their representations as graphs Γ_H with three permutations. In particular, for the Petersen graph these permutations written as products of cycles are given by $(01234)(57968)$, $(04321)(58697)$, $(05)(16)(27)(38)(49)$. Note that in this case the set H of single errors has an additional property: if $h \in H$ then $h^{-1} \in H$. We in fact showed that this additional property is guaranteed for any regular graph Γ of even degree. However, the following statement implies that this is not possible for the graph in Fig. 3.

Corollary 1: Any regular graph Γ of degree r is a graph Γ_H where $|H| = r$ and $h^{-1} \in H$ for any $h \in H$ if and only if r is even or r is odd and Γ has a 1-factor.

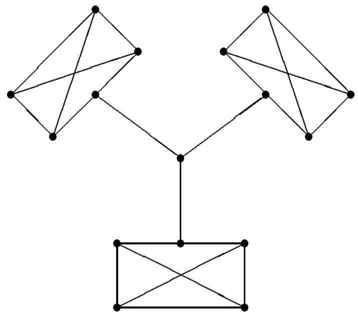

Fig. 3. A regular graph of degree 3 without 1-factors

C. Monotony on intersections of graphs with error metric

Now we prove a sufficient condition for a graph with error metric to be monotone on intersections. This allows one to find the value $N(V; t, d)$ for some types of single errors.

Lemma 6: Given a set H of permutations of a finite set V such that $H = H^{-1}$, the graph $\Gamma_H = \Gamma(V, E)$ with error metric $\rho(x, y)$ is monotone on intersections, if for any $h \in H$ and $x, y, z \in V$, the equalities $\rho(x, hy) = \rho(x, y) - 1$ and $\rho(hy, z) = \rho(y, z) + 1$ imply

(i) $\rho(hx, hy) = \rho(x, y)$,

(ii) $\rho(y, hz) = \rho(hy, z)$,

(iii) $\rho(x, hz) \leq \rho(x, z)$ if $\rho(x, z) = \rho(y, z)$.

Proof: Let for given $k \geq 2$ and $t = 1, ..., s$ the maximum $m_k(V; t)$ (see (47)) is attained at $x, y \in V$, $\rho(x, y) = k$. It is clear that there exists $h \in H$ such that $\rho(x, hy) = k-1$. To prove Lemma, it is sufficient to check that

$$\sum_{i=0}^{t}\sum_{j=0}^{t} p_{i,j}(x, hy) \geq \sum_{i=0}^{t}\sum_{j=0}^{t} p_{i,j}(x, y).$$

Let there exist $z \in V$ such that $z \in B_t(x) \bigcap B_t(y)$ but $z \notin B_t(x) \bigcap B_t(hy)$. It is possible only if $\rho(x, z) \leq t$, $\rho(y, z) = t$ and $\rho(hy, z) = t + 1$. Then consider hz and note that $\rho(hz, hy) = t$ by property (i), $\rho(y, hz) = t + 1$ by property (ii) and $\rho(x, hz) \leq t$ by property (iii). This means that $hz \in B_t(x) \bigcap B_t(hy)$ but $hz \notin B_t(x) \bigcap B_t(y)$ and completes the proof. □

Corollary 2: The Hamming graph $\Gamma_H = \Gamma(F_q^n, E)$ where H consists of all $(q-1)n$ substitutions is monotone on intersections and for any n, q, t, and d $(d \leq 2t)$,

$$N(F_q^n; t, d) = \sum_{i=0}^{t-\lceil \frac{d}{2} \rceil} \binom{n-d}{i} (q-1)^i$$

$$\times \sum_{k=d-t+i}^{t-i} \sum_{l=d-t+i}^{t-i} \binom{d}{k} \binom{d-k}{l} (q-2)^{d-k-l}.$$

In particular,

$$N(F_2^n; t, d) = \sum_{i=0}^{t-\lceil \frac{d}{2} \rceil} \binom{n-d}{i} \sum_{k=d-t+i}^{t-i} \binom{d}{k}$$

and

$$N(F_2^n; t, 2t) = N(F_2^n; t, 2t-1) = \binom{2t}{t}. \qquad (55)$$

Proof: Use Lemma 6 to prove that the Hamming graph is monotone on intersections. Let for a substitution $h_i^a \in H$ (see Example 2.1) $\rho(x, hy) = \rho(x, y) - 1$ and $\rho(hy, z) = \rho(y, z) + 1$. Then $x_i = y_i + a \bmod q$ and $z_i = y_i$, and the properties (i)-(iii) are held. Since the Hamming graph is distance-regular, by Lemma 3, for calculation of $N(F_q^n; t, d)$ we can find $|B_t(x) \bigcap B_t(y)|$ for any points $x, y \in F_q^n$ such that $\rho(x, y) = d$. Let $z \in F_q^n$ differs from x and y in i positions among those where x and y coincide and coincides with x in k positions and with y in l positions among those where x and y differ. If $z \in B_t(x) \bigcap B_t(y)$, then $0 \leq i \leq n - d$, $0 \leq k + l \leq d$, $d - k + i \leq t$, $d - l + i \leq t$ and hence $i \leq t - d/2$, $k \leq t - i$, $l \leq t - i$. This implies Corollary taking account of $\binom{a}{b} = 0$ if $b < 0$ or $b > a$. □

As an example, note that if $C \subseteq F_q^n$ is a code with the minimum distance $d = 3$, then the minimum number $M(C; t, 0) = N(F_q^n; t, 3) + 1$ of sequences that are sufficient to reconstruct an arbitrary point $x \in C$ for $t = 1$, $t = 2$, and $t = 3$ errors is equal to 1, $6(q-1)+1$ and $6(n-3)(q-1)^2 + q^3 + 1$, respectively. Note also that, for a code $C \subseteq F_q^n$ with the minimum distance $d = 2t - 1$, the equality (55) means that the minimum number of sequences $M(C; t, 0) = N(F_q^n; t, 2t-1) + 1$ that are sufficient to reconstruct an arbitrary point $x \in C$ in the combinatorial channel with at most t substitutions in n successive symbols equals $\binom{2t}{t} + 1$ independent of length n.

Corollary 3: The Johnson graph $\Gamma_H = \Gamma(J_w^n, E)$ where H consists of all $\binom{n}{2}$ transpositions of two symbols is monotone on intersections and for any n, w, t and d,

$$N(J_w^n; t, d) = \sum_{i=0}^{t} \binom{n-w-d}{i}$$

$$\times \sum_{a=0}^{t-i}\sum_{b=0}^{t-i} \binom{d}{a}\binom{d}{b}\binom{w-d}{a+b+i-d}.$$

Proof: We use Lemma 6 to prove the monotony on intersections of Γ_H. Since every transposition is an automorphism of Γ_H and an involution, the properties (i) and (ii) are satisfied. Assuming $\rho(x, y) = d$, $\rho(x, hy) = d - 1$, and $\rho(hy, z) = \rho(y, z) + 1$, we shall show that $\rho(x, hz) \leq \rho(x, z)$. Then we shall apply Lemma 3 to calculate $N(J_w^n; t, d)$, since the Johnson graph is distance-regular. Let $I = \{1, ..., n\}$ and $I_x = \{j \in I : x_j = 1\}$ for any $x = (x_1, ..., x_n)$. Then $|I_x \bigcap I_y| = w - d$, $|I_x \bigcap (I \backslash I_y)| = |I_y \bigcap (I \backslash I_x)| = d$ and hence h transposes one element of $I_x \bigcap (I \backslash I_y)$ with an element of $I_y \bigcap (I \backslash I_x)$. Since this transposition h applied to z increases its distance from y, it must transpose one element of $I_z \bigcap I_y \bigcap (I \backslash I_x)$ with an element of $(I \backslash I_z) \bigcap (I \backslash I_y) \bigcap I_x$. However, in this case the action of h on z decreases its distance from x. In order to calculate $|B_t(x) \bigcap B_t(y)|$ put $|I_z \bigcap I_x \bigcap (I \backslash I_y)| = a$, $|I_z \bigcap I_y \bigcap (I \backslash I_x)| = b$, $|I_z \bigcap (I \backslash (I_x \bigcup I_y))| = i$ and note that $|I_z \bigcap I_x \bigcap I_y| = w - a - b - i$, $d(x, z) = i + b$, and $d(y, z) = i + a$. This implies the formula above. □

Note that (9) and (11) are special cases of Corollaries 2 and 3 for $d = 1$.

Lemma 7: There exist distance-transitive graphs that are not monotone on intersections.

In particular, the following example of such graph ([2], p. 363) was communicated to the author by A. Brouwer. Consider a graph Γ whose set V of vertices is formed by 2^{10} cosets of the punctured perfect Golay (23,12,7)-code (see [16]). Two vertices are adjacent if and only if the Hamming distance between the corresponding sets is one. The sought graph Γ' is the bipartite double of Γ. The graph Γ' is distance-transitive and has the diameter 7. However, $m_1(V'; 2) = 44 < m_2(V'; 2) = 74$.

IV. Probabilistic channels

The problem of the efficient reconstruction of an unknown sequence distorted by errors which occur with certain probabilities is reduced to an optimization problem of multiple transmission of an arbitrary sequence over a probabilistic channel (see Fig. 1). It should be taken into account that for discrete channels, in general, the exact reconstruction of a sequence is possible only with a certain probability. For the channels with continious input and output, in general, the probability of the exact reconstruction of a multiply transmitted sequence equals zero and we can only attempt to reconstruct this sequence with a certain accuracy. In this section, we consider the problem of finding the minimum number of transmissions of a sequence over a discrete memoryless channel sufficient to reconstruct the sequence exactly (or within a given Hamming distance) with a permissible error probability. We shall also consider a similar problem for continuous channels with discrete time and additive noise: the efficient reconstruction of an arbitrary real sequence within a given Euclidean distance. It should be noted that these problems can be treated as those of mathematical statistics for the known probability distribution. However, we shall see that some methods and results of theory of information transmission can be successfully applied to these problems.

A. Optimal N-recontructors for discrete channels

Consider the problem of reconstruction of an unknown sequence $x = (x_1, ..., x_n) \in F_q^n$ provided that the permissible errors transform x to vectors $y = (y_1, ..., y_n) \in F_r^n$ (of the same length) with some probabilities. The problem is to find the minimum number N such that an arbitrary $x \in F_q^n$ can be reconstructed with a preset accuracy and error probability from a set $\{y_1, ..., y_N\}$ of its distorted patterns.

In order to give the precise sense to this problem we use the Shannon notion of a discrete memoryless channel with input and output alphabets F_q and F_r. Such a channel is characterized by the property that each letter of the output sequence is statistically dependent only on the corresponding letter of the input sequence. The channel is specified by a *transition matrix* $C = (p_{i,j})$ of the size $q \times r$, where $p_{i,j} = P(j|i)$ is the probability of receiving the letter $j \in F_r$ when the letter $i \in F_q$ is transmitted; $\sum_{j \in F_r} P(j|i) = 1$ for any $i \in F_q$. We shall denote this channel (as its transition matrix) by C. Thus, the probability $P_C(y|x)$ of receiving $y = (y_1, ..., y_n) \in F_r^n$ when $x = (x_1, ..., x_n) \in F_q^n$ is transmitted over a discrete memoryless channel C can be expressed as follows:

$$P_C(y|x) = \prod_{k=1}^{n} P(y_k|x_k).$$

For any $x = (x_1, ..., x_n) \in F_q^n$ we can consider the sequence $(y_1, ..., y_N)$ of elements of F_r^n as a sequence of patterns of x distorted by errors in the channel C. As distinguished from the case of the exact reconstruction for combinatorial channels, some elements of this sequence might be identical. We again consider $y_j = (y_{1,j}, ..., y_{n,j}) \in F_r^n$, $j = 1, ..., N$, as columns of a matrix $Y = (y_{i,j})$ over F_r of the size $n \times N$. All elements of each row of Y are images of the same letter. Denote by $Y_{n,N}$ the set of all r^{nN} matrices Y over F_r of the size $n \times N$. Let $\mathcal{F}_N$ be the set of all mappings $F : Y_{n,N} \to F_q^n$, $n = 1, 2, \ldots$, which are referred to as *N-reconstructors*. For a fixed $x = (x_1, ..., x_n) \in F_q^n$ we can consider $Y = (y_{i,j})$, $Y \subseteq Y_{n,N}$, as a random variable with the probability assignment

$$P_C(Y|x) = \prod_{k=1}^{n} \prod_{j=1}^{N} P(y_{k,j}|x_k). \tag{56}$$

For any integer-valued function $d = d(n)$, $0 \le d < n$, and any $F \in \mathcal{F}_N$ one can calculate the error probability

$$P_C(F, x, d, N) = \sum_{Y \in Y_{n,N},\ d_H(F(Y),x) > d} P_C(Y|x) \tag{57}$$

of reconstructing $x \in F_q^n$ with at most d wrong letters (here $d_H(z, x)$ is the Hamming distance). Note that the case $d = 0$ corresponds to the exact reconstruction. We set

$$P_C(n, d, N) = \min_{F \in \mathcal{F}_N} \max_{x \in F_q^n} P_C(F, x, d, N) \tag{58}$$

and call an N-reconstructor F *optimal* if it gives the minimum in (58) for all n (we shall see that optimal N-reconstructors exist for any function $d = d(n)$). The function $P_C(n, d, N)$ is a non-increasing function in N because one can consider that $Y_{n,N} \subset Y_{n,N+1}$. For any discrete memoryless channel C, ε $(0 < \varepsilon < \frac{1}{2})$, and integers n and d $(0 \le d < n)$, denote by $N_C(n, d, \varepsilon)$ the minimum integer N such that $P_C(n, d, N) \le \varepsilon$ (we shall see that such integer N exists except some degenerate cases). Thus, $N_C(n, d, \varepsilon)$ is the minimum number of repeated transmissions that allow one to reconstruct any sequence of length n with accuracy up to d letters with the error probability at most ε. Our aim is to obtain bounds on $N_C(n, d, \varepsilon)$ which determine its asymptotic behavior when $n \to \infty$ under some restrictions to the functions $\varepsilon = \varepsilon(n)$ and $d = d(n)$.

B. Reducible N-recontructors for discrete channels

For a discrete memoryless channel C it is natural to consider a class of N-reconstructors $(N = 1, 2, \ldots)$ whose action reduces N-tuple transmission of a message over C to its unitary transmission over another "improved" memoryless channel C_N. An N-reconstructor for a memoryless channel C (of size $q \times r$) is called *reducible*, if there exists a memoryless channel C_N (of size $q \times q$) such that for any n $(n = 1, 2, \ldots)$ and $x, z \in F_q^n$,

$$\sum_{Y \in Y_{n,N},\ F(Y) = z} P_C(Y|x) = P_{C_N}(z|x).$$

It is worth pointing out that any partition D of F_r^N into q subsets D_i (decoding regions of $i \in F_q$) uniquely defines an N-reconstructor F^D as follows: if $Y = (y_{i,j})$, $Y \subseteq Y_{n,N}$, $n = 1, 2, \ldots$, and $(y_{i,1}, ..., y_{i,N}) \in D_{z_i}$, $i = 1, ..., n$, then

$$F^D(Y) = (z_1, ..., z_n).$$

Since

$$\sum_{Y \in Y_{n,N},\ F^D(Y) = z} P_C(Y|x) = \prod_{i=1}^{n} P_C(D_{z_i}|x_i),$$

F^D is a reducible N-reconstructor for any channel C and the the corresponding transition matrix $C_N = C_{N,D}$ has entries $p_{i,j}^{N,D} = P_C(D_j|i)$, $i, j \in F_q$. On the other hand, any reducible N-reconstructor F for a channel C generates a partition $D = \bigcup_{j=0}^{q-1} D_j$ of F_r^N (or of some its subset) where $D_j = \{y : y \in F_r^N, F(y) = j\}$. In general, F differs from F^D for this partition. However, the corresponding channels C_N and $C_{N,D}$ (and hence the error probabilities (57)) must coincide for these N-reconstructors, because

$$P_{C_N}(j|i) = \sum_{Y \in Y_{1,N},\ F(Y)=j} P_C(Y|i) = P_C(D_j|i) = p_{i,j}^{N,D}.$$

This allows us to restrict our consideration of reducible N-reconstructors to N-reconstructors F^D where D is a partition F_r^N into q subsets.

Denote by $\mathcal{F}_N^\circ$ the set of all reducible N-reconstructors. By analogy with (58) we set

$$P_C^\circ(n,d,N) = \min_{F \in \mathcal{F}_N^\circ} \max_{x \in F_q^n} P_C(F,x,d,N) \tag{59}$$

and call a reducible N-reconstructor F (in particular, $F = F^D$) *optimal* if it gives the minimum in (59) for all n and d, $0 \le d < n$ (we shall see that optimal reducible N-reconstructors exist).

For a reducible N-reconstructor F^D, the value

$$1 - p_{i,i}^{N,D} = 1 - P_C(D_i|i)$$

characterizes the error probability of transmitting a letter $i \in F_q$ over this channel $C_{N,D}$ or of its recovery at the output of the N-reconstructor F^D. Let

$$P_C(N) = \min_D \max_{i \in F_q}(1 - P_C(D_i|i)) \tag{60}$$

where the minimum is taken over all partitions D of F_r^N into q decoding regions. It is evident that $P_C(N) = 1 - p_N$ where

$$p_N = \max_D \min_{i \in F_q} p_{i,i}^{N,D}. \tag{61}$$

Example 4.1: For the channel

$$C = \begin{pmatrix} 1-p & p \\ s & 1-s \end{pmatrix}, \tag{62}$$

$N = 2$, and a partition D of F_2^2 such that $D_0 = \{00, 01, 10\}$, $D_1 = \{11\}$ we have

$$C_{2,D} = \begin{pmatrix} 1-p^2 & p^2 \\ s\,(2-s) & (1-s)^2 \end{pmatrix}.$$

(For $p^2 = s\,(2-s)$ the channel $C_{2,D}$ is a symmetric one !) Let $0 < s \le p \le 1/2$ and $p^2 \le s\,(2-s)$. Then in the case $s\,(2-s) \le p$ this partition and $i = 1$ provide the extrema in (60) and (61) and imply that $P_C(2) = s\,(2-s)$. In the case $s\,(2-s) \ge p$ we have $P_C(2) = p$ and this is attained on the partition $D_0 = \{00, 01\}$, $D_1 = \{10, 11\}$ for $i = 0$.

Now we prove the existence of optimal reducible N-reconstructors and find their error probability (59). For any integers n and d, $0 \le d \le n$, and any real vector $u = (u_1, ..., u_n)$ where $0 \le u_i \le 1$, $i = 1, ..., n$, consider the function $g(n,d,u) = g(n,d,u_1,...,u_n)$ defined by the following expression:

$$g(n,d,u_1,...,u_n) = \sum_{m=d+1}^{n} \sum_{I \subseteq I_n, |I|=m} \prod_{i \in I} u_i \prod_{j \in I_n \setminus I} (1-u_j),$$

where $I_n = \{1, ..., n\}$. One can check that for $0 \le d < n$,

$$\frac{\partial g(n,d,u)}{\partial u_h} = \sum_{I \subseteq I_n \setminus \{h\}, |I|=d} \prod_{i \in I} u_i \prod_{j \in I_n \setminus \{I \bigcup \{h\}\}} (1-u_j).$$

Therefore,

$$g(n,d,u_1,...,u_n) \le S(n,d,\max_{1 \le i \le n} u_i) \tag{63}$$

where

$$S(n,d,p) = \sum_{i=d+1}^{n} \binom{n}{i} p^i \,(1-p)^{n-i}.$$

We shall also use the fact that $S(n,d,p)$ grows in p, $0 \le p \le 1$.

Lemma 8: For any integers n, d, and N ($0 \le d < n$, $N \ge 1$),

$$P_C^\circ(n,d,N) = S(n,d,P_C(N)) \tag{64}$$

Proof: For any (reducible) N-reconstructor F^D we have

$$P_C(F^D,x,d,N) = g(n,d,u_1,...,u_n)$$

where $u_i = 1 - P_C(D_{x_i}|x_i)$, $i = 1, ..., n$. According to (60) and (63) this implies that

$$P_C(F^D,x,d,N) \le S(n,d,P_C(N)). \tag{65}$$

If one considers the partition D of F_r^N and the letter $i_0 \in F_q$ for which the extrema in (60) are attained, then for $x = i_0^n$ the equality in (65) holds. Using the fact that the minimum in (59) is attained at an N-reconstructor F^D, we get (64). □

At the first sight it might seem that an optimal reducible N-reconstructor must be optimal in the class of all N-reconstructors since we consider memoryless channels C. However, this is not the case. The matter is that for a fixed n any (in particular, optimal) N-reconstructor F is uniquely defined by the following partition $D^{(n)}$ of the set $Y_{n,N}$ of all r^{nN} matrices Y over F_r of the size $n \times N$ into q^n decoding regions $D_z^{(n)}$, $z \in F_q^n$:

$$F(Y) = z \text{ if and only if } Y \in D_z^{(n)}. \tag{66}$$

Since $P_C(F,x,d,N)$ (see (57)) for $x \in F_q^n$ depends only on the partition $D^{(n)}$, one can find an optimal N-reconstructor F choosing for any n, $n = 1, 2, \ldots$, and $d = d(n)$ a partition $D^{(n)}$ which minimizes $\max_{x \in F_q^n} P_C(F,x,d,N)$. Note that for a reducible N-reconstructor F^D, the corresponding partition $D^{(n)}$ of $Y_{n,N}$ (which we shall denote by D^n) is obtained from the partition D of F_r^N as follows: $Y = (y_{i,j}) \in D_z^n$ where $z = (z_1, ..., z_n)$ if and only if $(y_{i,1}, ..., y_{i,N}) \in D_{z_i}$ for every $i = 1, ..., n$.

Example 4.1 (continued): The reducible 2-reconstructor F^D considered in Example 4.1 for $n = 2$ implies the following partition D^2:

$$D_{00}^2 = \left\{ \begin{matrix}00\\00\end{matrix}, \begin{matrix}00\\01\end{matrix}, \begin{matrix}00\\10\end{matrix}, \begin{matrix}01\\00\end{matrix}, \begin{matrix}01\\01\end{matrix}, \begin{matrix}01\\10\end{matrix}, \begin{matrix}10\\00\end{matrix}, \begin{matrix}10\\01\end{matrix}, \begin{matrix}10\\10\end{matrix} \right\}, D_{11}^2 = \left\{ \begin{matrix}11\\11\end{matrix} \right\},$$

$$D_{01}^2 = \left\{ \begin{matrix}00\\11\end{matrix}, \begin{matrix}01\\11\end{matrix}, \begin{matrix}10\\11\end{matrix} \right\}, D_{10}^2 = \left\{ \begin{matrix}11\\00\end{matrix}, \begin{matrix}11\\01\end{matrix}, \begin{matrix}11\\10\end{matrix} \right\}.$$

For $0 < s \le p \le 1/2$ and $p^2 \le s(2-s) \le p$, F^D is an optimal reducible 2-reconstructor and, in particular,

$$P_C^\circ(2,0,2) = 1 - \min_{x \in F_2^2} P_C(D_x^2|x).$$

Since $P_C(D_{00}^2|00) = (1-p^2)^2$, $P_C(D_{01}^2|01) = P_C(D_{10}^2|10) = (1-p^2)(1-s)^2$, $P_C(D_{11}^2|11) = (1-s)^4$, we have $P_C^\circ(2,0,2) = 1-(1-s)^4$. On the other hand, carrying the last matrix from D_{00}^2 to D_{11}^2 we get another partition $D^{(2)}$ for which

$$\min_{x \in F_2^2} P_C(D_x^{(2)}|x) > (1-s)^4$$

if, in addition, $(1-p^2)^2 - p^2(1-p)^2 > (1-s)^4$. All required inequalities are held, for example, when $s = p^2$, and hence $P_C(2,0,2) < P_C^\circ(2,0,2)$ in this case.

Thus, $P_C(n,d,N)$ is, in general, smaller than $P_C^\circ(n,d,N)$. To estimate $P_C(n,d,N)$ from below we consider partitions D^* of F_r^N into q subsets D_i^* (decoding regions of i) satisfying the property of the *maximum likelihood*: if $y \in D_i^*$ then $P_C(y|i) \ge P_C(y|k)$ for any $k \in F_q$. Note that such a partition D^* is in general not uniquely defined, however, the value

$$\overline{P}_C(N) = \frac{1}{q}\sum_{i=0}^{q-1}(1 - P_C(D_i^*|i)) \tag{67}$$

does not depend on D^* and minimizes

$$\frac{1}{q}\sum_{i=0}^{q-1}(1 - P_C(D_i|i))$$

in the class of all partitions D of F_r^N into q subsets D_i. Therefore (see (60)),

$$\overline{P}_C(N) \le P_C(N) \le P_C^*(N) \tag{68}$$

where

$$P_C^*(N) = \min_{D^*}\max_{i \in F_q}(1 - P_C(D_i^*|i)) \tag{69}$$

and the minimum is taken over all partitions D^* of F_r^N into decoding regions of i, $i \in F_q$, satisfying the property of the maximum likelihood.

Example 4.1 (continued): For the channel (62) with $0 < s < p \le 1/2$ the partition D of F_2^2 above is the unique one satisfying this property. Therefore, using calculations above we get

$$\overline{P}_C(2) = \frac{p^2 + 2s - s^2}{2} < P_C(2) = p < P_C^*(2) = 2s - s^2$$

if $s(2-s) > p$.

Lemma 9: For any integers n, d, and N $(0 \le d < n$, $N \ge 1)$,

$$S(n,d,\overline{P}_C(N)) \le P_C(n,d,N) \le S(n,d,P_C^*(N)).$$

Proof: The upper bound follows from (64) and (68). To obtain the lower bound we consider for an N-reconstructor F the partition $D^{(n)}$ defined by (66) and note that

$$\max_{x \in F_q^n} \sum_{Y \in Y_{n,N},\ d_H(F(Y),x) > d} P_C(Y|x) \ge$$

$$\frac{1}{q^n}\sum_{x \in F_q^n}\ \sum_{Y \in Y_{n,N},\ d_H(F(Y),x) > d} P_C(Y|x). \tag{70}$$

With the notation $K_d(z) = \{x \in F_q^n : d_H(z,x) > d\}$, we can rewrite (70) as

$$\frac{1}{q^n}\sum_{Y \in Y_{n,N}}\ \sum_{x \in K_d(F(Y))} P_C(Y|x). \tag{71}$$

We shall prove that (71) attains a minimum value when F is a reducible N-reconstructor F^{D^*} (or equivalently when $D^{(n)} = (D^*)^n$), where D^* is a partition of F_r^N into q regions satisfying the property of the maximum likelihood. Then we shall find this minimum value. To this end we fix a matrix $Y \in Y_{n,N}$, denote by $u_1, ..., u_n$ its rows, and set $F(Y) = (z_1, ..., z_n)$. Suppose that there exists i, $i = 1, ..., n$, such that $u_i \notin D_{z_i}^*$. To simplify notations we assume that $i = n$ and define $a \in F_q$ by the condition: $u_n \in D_a^*$. We change the partition $D^{(n)}$ carrying the matrix Y from $D_{F(Y)}^{(n)}$ to $D_z^{(n)}$ where $z = (z_1, ..., z_{n-1}, a)$, and show that this can only decrease the sum (71). Indeed, this transformation does not change the contribution to the sum (71) of all $x = (x_1, ..., x_n)$ such that $x_n \ne z_n$ and $x_n \ne a$. Let $x = (x_1, ..., x_{n-1}, z_n)$ and $x' = (x_1, ..., x_{n-1}, a)$. If $d_H(F(Y),x) > d$ or $d_H(F(Y),x) < d$, then x and x' belong or do not belong to both $K_d(F(Y))$ and $K_d(z)$, and their total contribution is also not changed. In case $d_H(F(Y),x) = d$ we have $x \notin K_d(F(Y))$, $x' \in K_d(F(Y))$, $x \in K_d(z)$, $x' \notin K_d(z)$, and $P_C(Y|x')$ in (71) is replaced by $P_C(Y|x) = P_C(Y|x')P_C(u_n|z_n)/P_C(u_n|a)$ which is not larger by the property of the maximum likelihood. Repeating this process a sufficient number of times we verify that (71) takes the minimum value at $D^{(n)} = (D^*)^n$. The fact that (70) for $F = F^{D^*}$ equals $S(n,d,\overline{P}_C(N))$ can be easily proved by induction on n for all d, $d = 0, 1, ..., n$. For $n = 1$, $d = 0$ this is equivalent to the definition (67). For the induction step one can use the following recurrence

$$S(n+1,d,p) = (1-p)S(n,d,p) + pS(n,d-1,p).$$

and the fact that $S(n,-1,p) = 1$. This implies the lower bound on $P_C(n,d,N)$. □

C. Bounds on the minimum number of repeated transmissions

Now we are ready to obtain an asymptotically tight bound on the minimum number of repeated transmissions $N_C(n,d,\varepsilon)$ that allow one to reconstruct any sequence of length n with accuracy up to d letters with the error probability at most ε in a discrete memoryless channel C.

Enumerate rows and columns of the transition matrix $C = (p_{i,j})$ by letters of the alphabets F_q and F_r. For any $i, k \in F_q$, denote by $A(i,k)$ the subset (which may be empty) consisting of all $j \in F_r$ such that $p_{i,j}p_{k,j} = P(j|i)P(j|k) > 0$. For any s, $0 \le s \le 1$, let

$$\alpha_{i,k}(s) = \sum_{j \in A(i,k)} (P(j|i))^{1-s} (P(j|k))^s \tag{72}$$

and

$$\alpha(C) = \max_{i,k \in F_q,\ i \neq k} \min_{0 \le s \le 1} \alpha_{i,k}(s). \tag{73}$$

It is clear that $0 \le \alpha(C) \le 1$. Note that $\alpha(C) = 0$ if and only if $A(i,k)$ is empty for all $i, k \in F_q$. This means that any column of the transition matrix C contains at most one nonzero probability and, hence, each letter of any input sequence is uniquely defined by the corresponding letter of the output sequence. Note also that $\alpha(C) = 1$ if C contains two identical rows. The converse statement is true as well. Indeed, if for some distinct $i, k \in F_q$,

$$\min_{0 \le s \le 1} \sum_{j \in A(i,k)} (P(j|i))^{1-s} (P(j|k))^s = 1,$$

then $\sum_{j \in A(i,k)} P(j|i) = \sum_{j \in A(i,k)} P(j|k) = 1$ and, hence, $P(j|i) = P(j|k) = 0$ for any $j \notin A(i,k)$. From the necessary condition of equality in the Holder inequality it follows that $P(j|i) = P(j|k)$ for all $j \in A(i,k)$ and, hence, C contains two identical rows. The existence of these identical rows in C implies that, for any $y \in F_r^n$, $P_C(y|i^n) = P_C(y|k^n)$ where $i^n = (i, ..., i) \in F_q^n$, $i \in F_q$. This means that at least one of the sequences i^n and k^n cannot be reconstructed with error probability $\varepsilon < \frac{1}{2}$. In order to exclude these trivial cases we shall consider the *non-degenerate* channels C for which the transition matrix C does not have two identical rows and contains a column with at least two nonzero probabilities. The arguments above show that

$$0 < \alpha(C) < 1.$$

if (and only if) the channel C is non-degenerate.

The following statement is derived from the celebrated result on bounds for the probability of decoding error for a code with two code words due Shannon, Gallager and Berlekamp ([21], Theorem 5). In fact, we apply their arguments to calculate the corresponding bounds for the special code $\{i^N : i \in F_q\}$. This explains rather a simple formulation of the statement.

Theorem 4: For any non-degenerate discrete memoryless channel C,

$$\frac{e^{-\beta(C)\sqrt{N}}}{2q} (\alpha(C))^N \le \overline{P}_C(N) \le P_C^*(N) \le (q-1)(\alpha(C))^N$$

where

$$\beta(C) = \sqrt{2} \min \max_{j \in A(i,k)} \left| \ln \frac{P(j|i)}{P(j|k)} \right|$$

and the minimum is taken over all $i, k \in F_q$, $i \neq k$, such that $\alpha(C) = \min_{0 \le s \le 1} \alpha_{i,k}(s)$.

Proof: Consider a partition D of F_r^N into q subsets D_i, $i \in F_q$, satisfying the property of maximum likelihood: if $y \in D_i$ then $P_C(y|i) \ge P_C(y|k)$ for any $k \in F_q$ ($P_C(y|i)$ is defined by (56), $n = 1$). For any $i, k \in F_q$, $k \neq i$, denote by $A_{i,k}^N$ the set (may be empty) of all $y = (y_1, ..., y_N) \in F_r^N$ such that $P_C(y|i)P_C(y|k) > 0$. Since the channel C is non-generate, there exists $i \in F_q$ with $P_C(D_i|i) < 1$. For any such i the set A_i of all $k \in F_q$, for which the set $A(i,k)$ (and hence $A_{i,k}^N$) is not empty, is not empty as well. Moreover, for any $k \in A_i$, $A_{i,k}^N$ is the set of all words of length N over the alphabet $A(i,k)$ and

$$(\alpha_{i,k}(s))^N = \sum_{y \in A_{i,k}^N} (P_C(y|i))^{1-s} (P_C(y|k))^s .$$

Therefore, using the property of the maximum likelihood, for any numbers s_k, $0 \le s_k \le 1$, $k \in A_i$, we have

$$1 - P_C(D_i|i) = \sum_{k \in A_i} \sum_{y \in D_k \bigcap A_{i,k}^N} P_C(y|i)$$

$$\le \sum_{k \in A_i} \sum_{y \in A_{i,k}^N} (P_C(y|i))^{1-s_k} (P_C(y|k))^{s_k} = \sum_{k \in A_i} (\alpha_{i,k}(s_k))^N .$$

Choosing s_k, $k \in A_i$, so as the minimum of $\alpha_{i,k}(s)$ for $0 \le s \le 1$ to be attained at $s = s_k$ and using the definitions (69) and (73) we obtain the desirable upper bound.

Now fix $i, k \in A$, $k \neq i$, such that

$$\alpha(C) = \min_{0 \le s \le 1} \alpha_{i,k}(s) \text{ and } \beta(C) = \sqrt{2} \max_{j \in A(i,k)} \left| \ln \frac{P(j|i)}{P(j|k)} \right| .$$

Let the minimum of $\alpha_{i,k}(s)$ for $0 \le s \le 1$ be attained at $s = s^\star$ and hence

$$\alpha_{i,k}(s^\star) = \alpha(C). \tag{74}$$

Since D_i and D_k are disjoint, (67) yields

$$q\overline{P}_C(N) \ge \sum_{y \in F_r^N \setminus D_i} P_C(y|i) + \sum_{y \in F_r^N \setminus D_k} P_C(y|k)$$

$$\ge \sum_{y \in F_r^N} \min(P_C(y|i), P_C(y|k)) . \tag{75}$$

Taking into account that $\alpha(C) = \min_{0 \le s \le 1} \alpha_{i,k}(s) > 0$ and hence $\alpha_{i,k}(s) > 0$ for $0 \le s \le 1$, set

$$\mu(s) = N \ln \alpha_{i,k}(s) \tag{76}$$

and note that $\mu(s)$ is differentiable any number of times and

$$\mu(s) = \ln \sum_{y \in A_{i,k}^N} (P_C(y|k))^{1-s} (P_C(y|i))^s . \tag{77}$$

The following remarkable fact was established in [21]. If, for any s, $0 \le s \le 1$, one considers the log likelihood ratio

$$\Delta(y) = \ln \frac{P_C(y|k)}{P_C(y|i)},\ y \in A_{i,k}^N, \tag{78}$$

to be a random variable with probability assignment

$$Q_s(y) = \frac{(P_C(y|i))^{1-s} (P_C(y|k))^s}{\sum_{z \in A_{i,k}^N} (P_C(z|i))^{1-s} (P_C(z|k))^s}, \tag{79}$$

then the derivatives $\mu'(s)$ and $\mu''(s)$ are the mean and variance of $\Delta(y)$ respectively! In particular, this implies that for the set

$$G_s = \{y \in A_{i,k}^N : |\Delta(y) - \mu'(s)| \le \sqrt{2\mu''(s)}\},$$

we have

$$\sum_{y \in G_s} Q_s(y) \ge \frac{1}{2}, \tag{80}$$

by the Chebyshev inequality. Note also that

$$0 \le \mu''(s) = N\frac{\alpha''_{i,k}(s)}{\alpha_{i,k}(s)} - N\left(\frac{\alpha'_{i,k}(s)}{\alpha_{i,k}(s)}\right)^2 \le \frac{N}{2}\beta^2(C). \tag{81}$$

From (77)-(79) it follows that

$$P_C(y|i) = e^{\mu(s) - s\Delta(y)}Q_s(y),$$

$$P_C(y|k) = e^{\mu(s) + (1-s)\Delta(y)}Q_s(y),$$

and hence for any $y \in G_s$,

$$P_C(y|i) \ge e^{\mu(s) - s\mu'(s) - s\sqrt{2\mu''(s)}}Q_s(y),$$

$$P_C(y|k) \ge e^{\mu(s) + (1-s)\mu'(s) - (1-s)\sqrt{2\mu''(s)}}Q_s(y).$$

According to (75) and (80), to complete the proof it is sufficient to verify that this implies that for any $y \in G_{s^\star}$,

$$\min\left(P_C(y|i), P_C(y|k)\right) \ge (\alpha(C))^N e^{-\beta(C)\sqrt{N}}Q_{s^\star}(y). \tag{82}$$

If $\mu'(s)$ changes the sign for $0 \le s \le 1$, then $\mu'(s^\star) = 0$ and (82) holds due to (74), (76), and (81). If $\mu'(s) \ge 0$ or $\mu'(s) \le 0$ for $0 \le s \le 1$, then $s^\star = 0$ or $s^\star = 1$ respectively and (82) also holds in the both cases. □

Example 4.1 (continued): For the channel (62) with $0 < s \le p \le 1/2$

$$\alpha(C) = p\left(\frac{1-s}{p}\right)^\tau \left(\ln\frac{(1-p)(1-s)}{ps}\right) / \ln\frac{1-p}{s}$$

where

$$\tau = \left(\ln\frac{(1-p)\ln\frac{1-p}{s}}{p\ln\frac{1-s}{p}}\right) / \ln\frac{(1-p)(1-s)}{ps},$$

and $\beta(C) = \sqrt{2}\ln\frac{1-p}{s}$. In particular, $\alpha(C) = 2\sqrt{p(1-p)}$ when $s = p$.

Example 4.2 (continued): For the channel

$$C = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{pmatrix},$$

$\alpha(C) = \max\left(\frac{3}{4}, \frac{2\sqrt{2}+1}{4}\right) = \frac{2\sqrt{2}+1}{4}$ and $\beta(C) = \sqrt{2}\ln 2$. Note, that in this case $P_C(1^n|1^n) = 0$ for any n since $p_{1,1} = 0$. However, in similar cases one can also recover any sequence with a given error probability using a sufficient number N of its transmissions over the channel.

Theorem 5: Let $\varepsilon = \varepsilon(n) > 0$ and $d = d(n) \ge 0$ are functions such that $\varepsilon \to 0$ and $d/n \to 0$ as $n \to \infty$. Then for any non-degenerate discrete memoryless channel C,

$$N_C(n,d,\varepsilon) \sim \frac{\ln\frac{n}{d+1} + \frac{1}{d+1}\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\alpha(C)}}. \tag{83}$$

Proof: From the definition of the number $N = N_C(n,d,\varepsilon)$ and Lemma 9 it follows that

$$S(n,d,\overline{P}_C(N)) \le P_C(n,d,N) \le \varepsilon$$
$$< P_C(n,d,N-1) \le S(n,d,P_C^*(N-1)). \tag{84}$$

Denote by $x(n) = x(n,d,\varepsilon)$ the unique solution of the equation $S(n,d,x) = \varepsilon$ when $x \in [0,1]$. Since $\varepsilon \to 0$ we have

$$x(n) \lesssim \frac{d(n)}{n}$$

and hence $x(n) \to 0$ as $n \to \infty$. By monotony of $S(n,d,p)$ on p,

$$\overline{P}_C(N) \le x(n) < P_C^*(N-1).$$

Therefore Theorem 4 implies that $N \to \infty$ and

$$-\ln x(n) \sim -N\ln\alpha(C) \tag{85}$$

at the asymptotic process under consideration. By Bonferroni's inequality

$$\binom{n}{d+1}x^{d+1}(1-x)^{n-d-1} \le S(n,d,x) \le \binom{n}{d+1}x^{d+1}.$$

The use of Stirling's inequalities shows that

$$-\frac{1}{d+1}\ln S(n,d,x) = \ln\frac{d+1}{nx} + O(1)$$

if $d/n \to 0$ and $x \lesssim d/n$ as $n \to \infty$. On account of (85), this completes the proof. □

In particular, by Theorem 5 $N_C(n,d,\varepsilon)$ grows linearly with length n when the permissible error probability ε of reconstruction of a sequence with at most d wrong letters decreases exponentially with n. It is interesting to compare this with the following result which shows that $N_C(n,d,\varepsilon)$ is restricted when reconstruction of a sequence is permissible with a *fixed part* d/n of wrong letters and $\varepsilon = \varepsilon(n)$ decreases exponentially with n. We use the Chernoff bound

$$S(n,d,p) \le 2^{-T(\delta,p)n} \text{ if } \delta = \frac{d+1}{n} \ge p, \tag{86}$$

where

$$T(\delta,x) = -\delta\log_2\frac{x}{\delta} - (1-\delta)\log_2\frac{1-x}{1-\delta}.$$

For a fixed δ, $0 < \delta < 1$, $T(\delta,x)$ decreases from ∞ to 0 when x runs the interval $[0,\delta]$. For any $c > 0$ denote by $\gamma(\delta,c)$ the unique (in $[0,\delta]$) root of the equation $T(\delta,x) = c$.

Theorem 6: If $d+1 \ge \delta n$, $0 < \delta < 1$, and $\varepsilon = 2^{-cn}, c > 0$, then for any non-degenerate discrete memoryless channel C and any n, $n = 1, 2, \dots$,

$$N_C(n,d,\varepsilon) \le \left\lceil \frac{\ln\frac{q-1}{\gamma(\delta,c)}}{\ln\frac{1}{\alpha(C)}} \right\rceil. \tag{87}$$

Proof: Denote by N the right-hand side of (87). Using Theorem 4 we get

$$P_C^*(N) \leq (q-1)\,(\alpha(C))^N \leq \gamma(\delta, c) \leq \delta$$

and hence by Lemma 9 for $p = P_C^*(N)$, any n and $d+1 \geq \delta n$,

$$P_C(n,d,N) \leq S(n,d,p) \leq 2^{-T(\delta,p)n} \leq 2^{-T(\delta,\gamma(\delta,c))n} = 2^{-cn}.$$

□

Note that the Chernoff bound (86) shows that the probability of the event that there exist at least δn wrong letters in a sequence of length n, (unitary) transmitted over symmetric binary channel with parameter p, decreases exponentially with n when $\delta > p$. If it is desirable to have a better constant in the exponent or an exponent in the case $\delta \leq p$, one can use repeated transmissions and estimate the minimum number of necessary transmissions with the help of Theorem 6. In particular, in the case $p = 0.02$, $\delta = 0.01$, and $c = 0.1$ we get that five repetitions are sufficient independently of length n.

D. *Reconstruction for continuous channels with additive noise*

Together with discrete channels without memory we consider channels C with discrete time and an additive noise for which the input and output alphabet is the set R of all reals. We again assume that each letter of the output sequence is statistically dependent only on the corresponding letter of the input sequence and is the sum of the initial letter with a noise which is a continuous random variable ξ with mean 0. For simplicity we assume that the distribution function of ξ has a symmetric density $p(x) = p(-x)$. (In this section we prefer to use letters x with indices as variables of integration.) For any input sequence $\theta = (\theta_1, ..., \theta_n) \in R^n$, we can consider a sequence $(y_1, ..., y_N)$ of elements of R^n as a sequence of patterns of θ distorted by errors in the channel C. The sequence $(y_1, ..., y_N)$ can be represented as a matrix $Y = (y_{i,j})$ over R of the size $n \times N$ with columns $y_1, ..., y_N$, where $y_j = (y_{1,j}, ..., y_{n,j})$, $j = 1, ..., N$. Now we denote by $Y_{n,N}$ the set of all real matrices Y of the size $n \times N$ and by $\mathcal{F}_\mathcal{N}$ the set of all N-reconstructors F which give a continuous function $F : Y_{n,N} \to R^n$ for any $n = 1, 2, \ldots$. For a fixed $\theta = (\theta_1, ..., \theta_n) \in R^n$ we can consider $Y = (y_{i,j}) \subseteq Y_{n,N}$ as an nN-dimensional random variable with the density $\prod_{i=1}^n \prod_{j=1}^N p(x_{i,j} - \theta_i)$. For any $F \in \mathcal{F}_\mathcal{N}$ and $\delta > 0$ we can calculate the probability $Pr(||F(Y) - \theta|| > \delta)$ of the event that θ is reconstructed to a sequence at the Euclidean distance more than δ from θ. (In the sequel, $||z||$ stands for the Euclidean norm of $z \in R^n$.) The value

$$P_C(n,\delta,N) = \inf_{F\in\mathcal{F}_\mathcal{N}} \sup_{\theta\in R^n} Pr(||F(Y) - \theta|| > \delta) \tag{88}$$

characterizes the maximum error probability for the best reconstructing mapping within Euclidean distance δ. We call an N-reconstructor F *optimal* if the infimum in (88) is attained at this F for all n and δ. For any n ($n = 1, 2, ...$), δ ($\delta > 0$) and ε ($0 < \varepsilon < \frac{1}{2}$), denote by $N_C(n, \delta, \varepsilon)$ the minimum integer N such that $P_C(n, \delta, N) \leq \varepsilon$.

The problem of finding an optimal N-reconstructor is the classical statistical problem to find the minimax estimate F of an unknown vector $\theta \in R^n$ based on N observations $y_1, ..., y_N$ of a random value $y \in R^n$ with density $p(x - \theta)$ provided that the loss function $L(F(Y), \theta) = w(||F(Y) - \theta||)$ with

$$w(z) = \begin{cases} 0 & \text{if } 0 \leq z \leq \delta, \\ 1 & \text{if } z > \delta, \end{cases} \tag{89}$$

is used. This estimate F gives the infimum in the expression

$$\inf_F \sup_{\theta\in R^n} E_\theta L(F(Y), \theta)$$

where E_θ is the symbol of mathematical expectation for a fixed θ. We should mention known results in general outline (see for instance [11], [23]). Under some restrictions on the functions $p(x)$ and $w(z)$, there exists an unique vector $F^* = F^*(Y)$ which minimizes the integral

$$\int_{R^n} w(||F^* - \theta||) \prod_{i=1}^n \prod_{j=1}^N p(y_{i,j} - \theta_i) d\theta.$$

The estimate $F^*(Y)$ is called the Pitman estimate and is minimax in the class of estimates $F(Y)$ which are invariant with respect to shifts, i.e., $F(y_1 + z, ..., y_N + z) = F(y_1, ..., y_N) + z$ for any $z \in R^n$. Under some additional restrictions, the Pitman estimate $F^*(Y)$ is also minimax in the class of all estimates, by the Hunt-Stein theorem. If $p(x)$ is a normal density and $w(z)$ is defined by (89) (or $w(z) = z^2$), then all required conditions are held and

$$F^*(Y) = \frac{1}{N}\sum_{j=1}^N y_j. \tag{90}$$

Thus, in this case the estimate (90) is minimax for all $n \geq 1$ and $\delta > 0$. (The author disposes of a simple combinatorial proof of the fact.)

From now on we consider the Gaussian channel $C = G$ with variance σ^2 when ξ has the density

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \tag{91}$$

In this case for any n and σ, the N-reconstructor (90) is optimal and $P_G(n, \delta, N) = 1 - Pr(\|\overline{\xi}\|^2 \leq \delta^2)$ where $\overline{\xi} = \frac{1}{N}\sum_{j=1}^N \xi_j$ and $\xi_j = (\xi_{1,j}, ...\xi_{n,j})$. Each component $(\overline{\xi})_i$, $i = 1, ..., n$, of the vector $\overline{\xi}$ is the arithmetical average of N identical independent normally distributed variables with the mean 0 and variance σ^2 and, hence, is a normally distributed variable with the mean 0 and variance σ^2/N. Thus, similar to reducible N-reconstructors for a discrete memoryless channel, action of the optimal N-reconstructor (90) is equivalent to unitary transmission of a message $\theta = (\theta_1, ..., \theta_n) \in R^n$ over the same Gaussian channel with the decreased variance σ^2/N. Moreover, since $(\overline{\xi})_i$, $i = 1, ..., n$, are independent random variables, $\sum_{i=1}^{n} \left|(\overline{\xi})_i\right|^2$ has the χ^2 probability distribution

$$Pr(\sum_{i=1}^{n} \left|(\overline{\xi})_i\right|^2 \leq \delta^2) = \frac{1}{2^{n/2-1}\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{\delta\sqrt{N}}{\sigma}} x^{n-1} e^{-\frac{x^2}{2}} dx$$

and hence

$$P_G(n, \delta, N) = 1 - \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{\delta^2 N}{2\sigma^2}} \lambda^{\frac{n}{2}-1} e^{-\lambda} d\lambda. \qquad (92)$$

Lemma 10: For any even n, the number $N = N_G(n, \delta, \varepsilon)$ is uniquely defined from the following inequalities:

$$A(cN, \frac{n}{2} - 1) \leq \varepsilon < A(c(N-1), \frac{n}{2} - 1)$$

where $c = \frac{\delta^2}{2\sigma^2}$ and

$$A(\lambda, m) = e^{-\lambda} \sum_{i=0}^{m} \frac{\lambda^i}{i!}.$$

Proof: For any nonnegative integer m,

$$\frac{\partial A(\lambda, m)}{\partial \lambda} = -e^{-\lambda} \frac{\lambda^m}{m!}$$

and, hence, (92) implies that for even n

$$P_G(n, \delta, N) = A(cN, \frac{n}{2} - 1).$$

Since $P_G(n, \delta, N-1) > \varepsilon$, we get the statement of Lemma. □

As a numeric example note that for $n = 10$, any $\delta = 2\sigma$ and $\varepsilon = 0,002$, $N_G(n, \delta, \varepsilon) = 7$ by Lemma 10.

Theorem 7: If σ and δ are fixed, and $\varepsilon = \varepsilon(n) \to 0$ as $n \to \infty$, then

$$N_G(n, \delta, \varepsilon) \sim \begin{cases} \frac{\sigma^2}{\delta^2}(1 + \mu(\gamma))\, n & \text{if } \frac{-\ln \varepsilon}{n} \to \frac{\gamma}{2}\ (\gamma \geq 0), \\ 2\frac{\sigma^2}{\delta^2} \ln \frac{1}{\varepsilon} & \text{if } \frac{-\ln \varepsilon}{n} \to \infty, \end{cases}$$

where $\mu(\gamma)$ is the unique nonnegative solution of the equation $\mu - \ln(1 + \mu) = \gamma$.

Proof: Because $P_G(n, \delta, N) \leq P_G(n + 1, \delta, N)$ and hence $N_G(n + 1, \delta, \varepsilon) \geq N_G(n, \delta, \varepsilon)$, we can consider only even n. Put $m = n/2 - 1$. Since $A(\lambda, m)$ is a decreasing function in λ, by Lemma 10 the asymptotic behavior of $N_G(n, \delta, \varepsilon)$ coincides with that of $\lambda(n)/c$ where $\lambda(n)$ is the unique solution of the equation $A(\lambda, m) = \varepsilon(n)$. Find the asymptotic behavior of $A(\lambda, m)$ for $\lambda = m + x\sqrt{m}$, $x > 0$, as $n \to \infty$. Using ordinary arguments we get

$$\sum_{i=0}^{m} \frac{\lambda^i}{i!} \leq \frac{\lambda^m}{m!} \sum_{j=0}^{\infty} \left(\frac{m}{\lambda}\right)^j = \frac{\lambda^m}{m!} \frac{\lambda}{\lambda - m}$$

and

$$\sum_{i=0}^{m} \frac{\lambda^i}{i!} \geq \frac{\lambda^m}{m!} \sum_{j=0}^{\lfloor\sqrt{m}\rfloor - 1} \left(\frac{m - \sqrt{m}}{\lambda}\right)^j.$$

These inequalities and Stirling's formula show that $A(\lambda, m)$ is greater than a positive constant if x is a constant, and

$$A(\lambda, m) \sim \frac{1}{\sqrt{2\pi m}} (\frac{\sqrt{m}}{x} + 1) e^{-\sqrt{m}x + m\ln(1 + \frac{x}{\sqrt{m}})}$$

if $x \to \infty$. It follows that $x \to \infty$ if $A(\lambda, m) \to 0$. Moreover,

$$-\ln A(\lambda, m) \sim \frac{x^2}{2} \text{ if } x \to \infty \text{ and } \frac{x}{\sqrt{m}} \to 0; \qquad (93)$$

$$-\ln A(\lambda, m) \sim m\,(\mu - \ln(1 + \mu)) \text{ if } \frac{x}{\sqrt{m}} = \mu > 0; \quad (94)$$

$$-\ln A(\lambda, m) \sim x\sqrt{m} \text{ if } \frac{x}{\sqrt{m}} \to \infty. \qquad (95)$$

Let $x(n)$ be defined by $\lambda(n) = m + x(n)\sqrt{m}$. Then $x(n) \to \infty$ since $A(\lambda(n), m) = \varepsilon(n) \to 0$. Equations (93)-(95) allow us to find asymptotic behavior of $x(n)$ and $\lambda(n) = m + x(n)\sqrt{m}$ depending on that of $-\frac{1}{n} \ln \varepsilon(n) = -\frac{1}{n} \ln A(\lambda(n), m)$. □

In particular, $N_G(n, \delta, \varepsilon)$ grows linearly with sequence length n when the permissible error probability ε of reconstruction within a given Euclidean distance δ decreases exponentially with n.

V. Concluding remarks and open problems

The aim of this paper is to develop the theory of efficient reconstruction of sequences which deals with optimization problems of repeated transmission of information through combinatorial and probabilistic channels. There is a significant difference between these problems and the traditional problems of the theory of information transmission. The matter is that we consider repeatedly transmitting an arbitrary message in the noncoded form and minimize the number of retransmissions sufficient for reproducing the message with a preset accuracy and/or probability. This theory includes combinatorial, information-theoretical and statistics problems.

For combinatorial channels with types of single errors of considerable interest in coding theory, such as substitutions, transpositions, asymmetric errors, deletions and insertions, these optimization problems were solved. Moreover, simple algorithms for efficient reconstruction of sequences based on generalized threshold functions were found. However, there are numerous open problems connected with other types of single errors including their combinations, for example, substitutions, deletions and insertions. Interesting combinatorial problems also arise to find $N_H(V, t)$ for some subsets $V \subseteq F_q^n$, for example, for the set V of words with a given composition (in particular, for all permutations when $n = q$) in the case of transpositions.

The concept of the graph Γ_H with error metric develops the general construction of metrics on a finite or countable set V (in particular, $V \subseteq F_q^*$) introduced in [12]. At the first sight it is surprising that any finite graph Γ of degree r can be treated as a graph Γ_H with a set H of r single errors being permutations of vertices in the case of regular Γ. Coding theory, theory of sequences, and computational molecular biology (see [1], [6], [8], [18]) give numerous examples of types of single errors (partial one-to-one mappings of V) for which the property $H = H^{-1}$ or weaker parallelogram property are satisfied. An important problem is to describe types of single errors inherent to genes, genomes, and other objects of molecular biology and solve the considered problem on the minimum number of erroneous patterns sufficient for exact reconstruction. It is worth to mention a natural generalization of graphs Γ_H when each single error $h \in H$ has a positive weight and Γ_H is considered as a directed graph with weighted edges. In this case each ordered pair (x, z) of vertices of Γ_H is characterized by the "weighted distance" which equals the minimum sum of edge weights in a directed path jointing x with z. To advance the graph-theoretical approach it is significant to calculate $N(V; t, d)$ for some other graphs Γ_H and, in particular, to strengthen Lemma 6 for the Abelian graphs. An interesting problem is to find the minimum size $M(F_q^n; t, \rho)$ of a set Y in the Hamming metric ball $B_t(x)$ centered at an arbitrary $x \in F_q^n$ which allows one to approximate this x to within distance ρ and find the corresponding algorithm. In general, the majority algorithm is not suitable for $\rho > 0$.

Combinatorial channels essentially differ from probabilistic ones in that they admit of uniquely reproducing messages, whereas for probabilistic channels messages are reproduced with a certain probability. However, one should pay a large price for this possibility. A simple calculation for the combinatorial (n, t)-channel of Theorem 1 shows that if the number t of errors increases linearly with increasing the length n of messages, then uniquely reproducing messages requires that the number of different erroneous patterns increases exponentially with increasing n. On the other hand, by Theorem 5, for reconstructing any message of length n with a fixed probability $\varepsilon > 0$ (for instance, $\varepsilon = 10^{-9}$) at the output of a discrete probabilistic channel, a number of repetitions increasing logarithmically with increasing n is sufficient. (The difference is 2 orders of the magnitude!)

The notion of a reducible N-reconstructor for a discrete memoryless channel seems natural and fruitful. Although an optimal N-reconstructor is not in general reducible, bounds for reducible N-reconstructors were used for proof of the main Theorem 5, and the asymptotic (83) is also valid for the class of reducible N-reconstructors. Moreover, this notion gives rise to new extremum problems of transformation of discrete memoryless channels (as probabilistic sources) using repeated transmissions of a sequence.

The results on reproduction of a sequence for the Gaussian channel with the help of its repeated transmissions lie in the course of traditional problems of mathematical statistics. However, the question of the interest is the existence and construction of optimal N-reconstructors when the distribution function of noise has another symmetric density.

Acknowledgment

The author wishes to express his gratitude to A. Brouwer, Y. Golubev, and D. Hirschberg for comments and suggestions which helped to improve this paper.

References

[1] A. Apostolico and R. Giancarlo, "Sequence alignments and molecular biology", J. of Comp. Biology, vol. 5, no. 2, pp. 173–196, 1998.

[2] A.E. Brouwer, A.M. Cohen, A. Neumaier. Distance-Regular Graphs. Berlin: Springer Verlag, 1989.

[3] L. Calabi, "On the computation of Levenshtein's distances", TN-9-0030, Parke Math. Labs., Inc., Carlisle, Mass., 1967.

[4] L. Calabi and W.E. Hartnett, "Some general results of abstract coding theory with applications to the study of codes for the correction of synchronization errors", Inform. and Control, vol. 15, pp. 235–249, 1969.

[5] P. Delsarte, "An algebraic approach to the association schemes of coding theory", Philips Res. Repts. Suppl., vol. 10, 1973.

[6] D. Eppstein, Z. Galil, R. Giancarlo, G.F. Italiano, "Efficient algorithms for sequence analysis", in Methods in Communication, Security, and Computer Science, Springer Verlag, 1991.

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

[8] A. Guénoche, "Can we recover a sequence, just knowing all its subsequences of given length?", CABIOS, vol. 8, no. 6, pp. 569–574, 1992.

[9] F. Harary, Graph Theory, London: Addison-Wesley, 1969.

[10] D.S. Hirschberg, "Bounds on the number of string subsequences", in Proc. 10th Symp. on Combinatorial Pattern Matching, Warwick UK, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 1999.

[11] I.A. Ibragimov, R.Z. Hasminskii, Asymptotic theory of estimates, Moscow: Nauka, 1979 (in Russian). English translation: Statistical Estimation. Asymptotic Theory, Springer-Verlag, 1981.

[12] V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals, DAN, vol. 163, no. 4, pp. 845-848, 1965 (in Russian). English translation in Sov. Phys. -Dokl., vol. 10, no. 8, pp. 707-710, 1966.

[13] V.I. Levenshtein, "Elements of coding theory", in Discrete Mathematics and Math. Problems of Cybernetics, Nauka, Moscow, 1974, pp. 207–305 (in Russian). German translation in

[14] V.I. Levenshtein, "On perfect codes in deletion/insertion metric", Discretnaya Mathematika, vol. 3, no. 1, pp. 3-20, 1991 (in Russian). English translation in Discrete Math. Appl., vol. 2, no. 3, pp. 241–258, 1992.

[15] V.I. Levenshtein, "Reconstruction of objects from a minimum number of distorted patterns", DAN, vol. 354, no. 5, pp. 593-596, 1977 (in Russian). English translation in Doklady Mathematics, vol. 55, no. 3, pp. 417–420, 1997.

[16] F.J. MacWilliams, N.J.A. Sloane, The Theory of Error-Correcting Codes, New York: North-Holland, 1977.

[17] W.W. Peterson, E.J. Weldon, Jr., Error-Correcting Codes, The MIT Press, Cambridge, 1972.

[18] D. Sankoff and J.B. Kruskal, editors, Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley, Reading, MA, U.S.A., 1983.

[19] C. Shannon, "A mathematical theory of communication", Bell System Techn. J., vol. 27, pp. 379–423 and 623–656, 1948.

[20] C. Shannon, "A theorem on colouring the lines of a network", J. of Math. and Phys., vol. 28, pp. 148–151, 1949.

[21] C. Shannon, R.G. Gallager and E.R. Berlekamp, "Lower bounds to error probability for coding on discrete memoryless channels", Inform. and Control, vol. 10, pp. 65–103 and 522–552, 1967.

[22] V.G. Vizing, "On estimate of the chromatic class of a p-graph", Discretnyi Analiz, vol. 3, pp. 25-30, 1964 (in Russian). English translation in

[23] S. Zacks, The Theory of Statistical Inference. New York: Wiley, 1971.