Efficient Reconstruction of Sequences

Vladimir I. Levenshtein, *Associate Member, IEEE*

***Abstract*—In this paper, we introduce and solve some new problems of efficient reconstruction of an unknown sequence from its versions distorted by errors of a certain type. These erroneous versions are considered as outputs of repeated transmissions over a channel, either combinatorial channel defined by the maximum number of permissible errors of a given type, or a discrete memoryless channel. We are interested in the smallest N such that N erroneous versions always suffice to reconstruct a sequence of length n, either exactly or with a preset accuracy and/or with a given probability. We are also interested in simple reconstruction algorithms. Complete solutions for combinatorial channels with some types of errors of 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 nontrivial 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 Hamming distance d with error probability ε is found when d/n and ε tend 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*—Algorithms, combinatorial and probabilistic channels, error metric, graphs, reconstruction, repeated transmission, sequences.**

I. Introduction

TRADITIONAL problems of the 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. In the solution of these problems, a *coding* of the elements of V is used to introduce a redundancy to the messages so that the distance (in a certain metric) between the encoded messages would be sufficiently large and allow one to correct errors at the output of the channels. The efficiency of transmission is characterized by the amount of redundancy and by the complexity of the decoding algorithm. One method to combat errors is repeated transmission of a message, without coding. This is not efficient from the point of view of redundancy, but in various fields of science, such as informatics, molecular biology, and chemistry, there are situations when no other method is feasible. With that motivation we study in this paper the problem of recovering an unknown sequence (or message) $x = (x_1, \dots, x_n) \in V$ when a sufficiently large number of patterns (sequences) are known which are distorted versions of x. We can assume that the components of x belong to the alphabet $A_q = \{0, 1, \dots, q-1\}$, $q \ge 2$, and hence $V \subseteq A_q^n$.

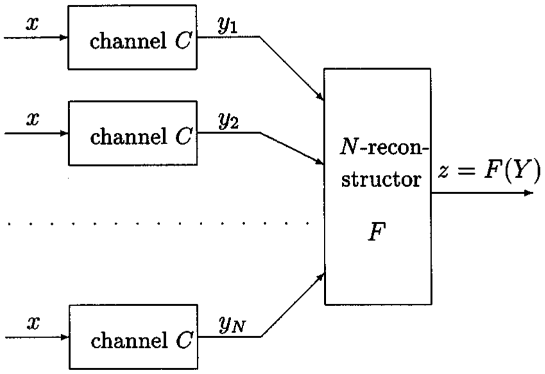

Fig. 1. Schematic diagram of reconstructing a sequence x.

For the precise formulation of the problem we can define a combinatorial or probabilistic channel C and assume that N patterns $y_1, \dots, y_N$ 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 a probabilistic channel C we consider an ordinary discrete noisy channel without memory [20], [7]. For restoring a sequence $x \in V$ in both cases we use an N-reconstructor F which maps the matrix Y formed by the columns $y_1, \dots, y_N$ to V. We call $F(Y)$ an *exact reconstruction* of x if $F(Y) = x$, and a *reconstruction of x within distance d* if $d_H(F(Y), x) \le d$ 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, we call the probability of the event $d_H(F(Y), x) > d$ the *error probability* of reconstructing x within distance d. In the case of a combinatorial channel C, we should certainly assume that all patterns $y_1, \dots, y_N$ are *different*, since otherwise exact reconstruction of x is not possible, for instance, when they coincide and differ from x. No such assumption is needed 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.

Manuscript received May 24, 1999; revised June 13, 2000. This work was supported by the Russian Foundation for Basic Research under Grant 98-01-00146. The material in this paper was presented in part at the IX All Union Conference on Coding Theory and Information Transmission, Odessa, Russia, 1988, at the 5th International Workshop on Algebraic and Combinatorial Coding Theory, Sozopol, Bulgaria, June 1–7, 1996, at the 2nd Siberian Congress on Applied and Industrial Mathematics, Novosibirsk, Russia, July 25–29, 1996, and at the IEEE International Symposiums on Information Theory, Ulm, Germany, July 29–July 4, 1997, and Sorrento, Italy, June 25–30, 2000.

The author is with the Keldysh Institute for Applied Mathematics, Russian Academy of Science, 125047 Moscow, Russia (e-mail: leven@spp.keldysh.ru).

Communicated by I. Csiszár, Associate Editor for Shannon Theory.
Publisher Item Identifier S 0018-9448(01)00465-5.

A natural measure of efficiency of a solution of the combinatorial problem under consideration is the minimum number N such that there exists an N-reconstructor which exactly reconstructs any $x \in V$ from 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 probabilistic 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 reconstructing 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 interest in coding theory. First we consider the combinatorial problem of reconstructing an unknown sequence $x = (x_1, \ldots, x_n) \in A_q^n$ when knowing N different sequences $y_1, y_2, \ldots, y_N \in A_q^n$ each of which differs from x in at most t components (i.e., obtained by at most t substitutions). What is the minimum number N which is sufficient to exactly reconstruct any $x \in A_q^n$? 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 $x \in A_2^5$ when $t = 2$? For any n, q, and t we show that this minimum number N equals $N(A_q^n, t) + 1$ where

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

and prove that in each row of the matrix Y formed by the columns $y_1, y_2, \ldots, y_N$, one letter of A_q occurs more often than others and it equals the component x_i of the unknown $x = (x_1, \ldots, x_n)$. 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 $x \in A_2^5$, 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 also be suitable.

The next combinatorial problem is to reconstruct an unknown sequence $x = (x_1, \ldots, x_n) \in A_2^n$ when knowing N different sequences $y_1, y_2, \ldots, y_N \in A_2^n$ 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 A_2^n consisting of all vectors of weight w. As an example, can we uniquely restore $x \in J_2^5$ from the matrix Y

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

of 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 the reconstruction of

$$x = (x_1, \ldots, x_n) \in J_w^n$$

from 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 ith row of Y is greater than $(N-1)w/n$ and $x_i = 0$ otherwise $(i = 1, \ldots, n)$. 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 the reconstruction of any $x \in A_q^n$ 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 the reconstruction of an arbitrary $x = (x_1, \ldots, x_n) \in A_q^n$ when knowing different sequences $y_1, y_2, \ldots, y_N \in A_q^{n-t}$, each obtained from x by deletions of exactly t symbols and hence is a subsequence $(x_{i_1}, \ldots, x_{i_{n-t}})$ of x, $1 \leq i_1 < \cdots < i_{n-t} \leq n$. If $N^-(A_q^n, t)$ denotes the maximum size of the set of common subsequences of length $n - t$ of two different sequences $x, z \in A_q^n$, then $N = N^-(A_q^n, t) + 1$ is the minimum number such that any $x \in A_q^n$ can be exactly reconstructed using N of its different subsequences of length $n - t$ (if they exist). We find $N^-(A_q^n, t)$ and prove that

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

with equality for $q = 2$. In particular, $N^-(A_2^7, 2) = 10$ and hence we can find a unique $x \in A_2^7$ such 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 $x \in A_q^n$ with the help of $N = N^-(A_q^n, t) + 1$ 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 $N^+(A_q^n, t)$ of common supersequences of length $n + t$ of two different sequences from A_q^n and present an algorithm for reconstruction of an arbitrary

$x \in A_q^n$ when knowing $N = N^+(A_q^n, t) + 1$ different supersequences of length $n + t$ (obtained from x by insertions of t symbols of A_q).

The considered combinatorial problems show that in fact we deal with the same problem for different metrics on A_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 when knowing a sufficient 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 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 $h(x) = y$. Then 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 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 the 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 N such that any N vertices in the metric ball of radius t around any vertex x suffice for exact reconstruction of x. The property of *monotonicity on intersections* is introduced and a sufficient condition so that a graph has this property is found. This property allows us to easily calculate the values $N(t, d)$ for some graphs.

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 in a code $C \subseteq V$ of minimum 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 $N(t, 2t-1) = \binom{2t}{t}$. 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, for the combinatorial channel with at most t substitutions of symbols, is equal to $\binom{2t}{t} + 1$ independently of the 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 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 reducible N-reconstructors which have a remarkable property: N-tuple transmission of a message over a discrete memoryless channel C using a reducible N-reconstructor (see Fig. 1) is equivalent to *single* 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 *nondegenerate* 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 introduced in [22] for finding the zero rate exponent and note that $0 < \alpha(C) < 1$ if and only if C is nondegenerate. 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 nondegenerate 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 not faster than an exponent in n, a bounded (i.e., independent of n) number of repetitions is 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 $x = (x_1, \ldots, x_n)$ within Euclidean distance δ with error probability ε, is reduced to a classical statistical problem for the minimax estimate. In the case of a Gaussian channel, 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 single transmission of x over the same channel with variance divided by N. We find the asymptotic behavior of $N(n, \delta, \varepsilon)$ for this channel.

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

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

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 x. 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, $x \in V$ can be reconstructed with the help of any N *different* elements of $B_t(x, H)$ if they exist. In this paper, we shall determine $N_H(V, t)$ for different $H, n, V \subseteq A_q^n$, and t.

Now we give examples of 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 A_q^n.

Example 2.1 (Substitutions):

$$H = \{h_i^a \colon a \in A_q \backslash \{0\},\ i = 1, 2, \ldots\}$$

where the single error h_i^a is defined on all $x = x_1 \ldots x_n \in A_q^n$ such that $n \geq i$ and replaces the letter x_i in x by the letter $x_i + a \mod q$. In this case, $|H_n| = (q-1)n$.

Example 2.2 (Asymmetric Errors):

$$H = \{h_i^{>} \colon i = 1, 2, \ldots\}$$

and

$$H = \{h_i^{<} \colon i = 1, 2, \ldots\}$$

where the single error $h_i^{>}$ $(h_i^{<})$ is defined on all $x = x_1 \cdots x_n \in A_q^n$ such that $n \geq i$ and $x_i < q-1$ $(x_i > 0)$ and replaces the letter x_i in x by the letter x_i+1 (x_i–1, respectively). In both cases, $|H_n| = n$.

Example 2.3 (Cyclic Errors):

$$H = \{h_i^{\pm} \colon i = 1, 2, \ldots\}$$

where the single error $h_i^{\pm}$ is defined on all $x = x_1 \cdots x_n \in A_q^n$, $q \geq 3$, such that $n \geq i$ and replace the letter x_i in x by the letter $x_i \pm 1 \mod q$. In this case $|H_n| = 2n$.

Example 2.4 (Transpositions):

$$H = \{h_{i,j} \colon i, j = 1, 2, \ldots, i < j\}$$

where the single error $h_{i,j}$ is defined on all $x = x_1 \cdots x_n \in A_q^n$ 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):

$$H = \{h_i^{a,-} \colon a \in A_q,\ i = 1, 2, \ldots\}$$

where the single error $h_i^{a,-}$ is defined on all $x = x_1 \cdots x_n \in A_q^n$ such that $n \geq i$, $x_i = a$, and maps x to the word $x_1 \cdots x_{i-1} x_{i+1} \cdots x_n$ of length $n-1$. In this case $|H_n| = qn$.

Example 2.6 (Insertions):

$$H = \{h_i^{a,+} \colon a \in A_q,\ i = 0, 1, \ldots\}$$

where the single error $h_i^{a,+}$ is defined on all $x = x_1 \cdots x_n \in A_q^n$ such that $n \geq i$ and maps x to the word $x_1 \cdots x_i a x_{i+1} \cdots x_n$ 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, [18]) also admit a similar description.

In addition to determining $N_H(V, t)$, it is also important to find a simple algorithm for reconstructing $x \in V$ from N different words $y_1, \ldots, y_N$ in $B_t(x, H)$ where $N = N_H(V, t)$+1. The following definitions will be needed for that purpose. The *composition* of a word $u \in A_q^n$ is

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

where $k_i(u)$ is the number of occurrences of the letter $i \in A_q$ in u. The *ordered composition* of u is

$$l(u) = (k_{\theta_0}(u), \ldots, k_{\theta_{q-1}}(u)) \qquad (4)$$

where

$$\theta(u) = (\theta_0, \theta_1, \ldots, \theta_{q-1}) \qquad (5)$$

is a permutation of $A_q = \{0, 1, \ldots, q-1\}$ such that

$$k_{\theta_0}(u) \geq k_{\theta_1}(u) \geq \cdots \geq k_{\theta_{q-1}}(u).$$

The last conditions need not uniquely determine the permutation $\theta(u)$ of A_q, but it will be uniquely defined if we additionally require that $\theta_i < \theta_j$ whenever $k_i(u) = k_j(u)$, $i < j$. With $\theta(u)$ as defined above, we define the *majority function* $m_q \colon A_q^N \rightarrow A_q$ by

$$m_q(u) = \theta_0. \qquad (6)$$

If a letter $a \in A_q$ occurs in u more often than any other letter then $m_q(u) = a$. We will also need the *threshold function* $f_\tau \colon R^m \rightarrow A_m$ (typically, m will be q or $t+1$). Given a vector of thresholds $\tau = (\tau_0, \tau_1, \ldots, \tau_{m-1}) \in R^m$, we define

$$f_\tau(w_0, w_1, \ldots, w_{m-1}) = \min\{i \in A_m \colon w_i > \tau_i\}. \qquad (7)$$

The threshold function $f_\tau(w)$ is well defined for all $w = (w_0, w_1, \ldots, w_{m-1}) \in R^m$ such that

$$\sum_{i=0}^{m-1} w_i > \sum_{i=0}^{m-1} \tau_i. \qquad (8)$$

Note that if $w_i > \tau_i$ holds for only one $i \in A_m$ then $f_\tau(w) = i$. Note also that the majority function (6) can be expressed as $f_{0,\ldots,0}(l(u))$ (with q zero thresholds). As an example, for $q = 3$ and $u = 0211012$ we have $k(u) = (2, 3, 2)$, $\theta(u) = (1, 0, 2)$, $l(u) = (3, 2, 2)$, $m_3(u) = 1$, $f_{3,2,1}(k(u)) = 1$, and $f_{3,2,1}(l(u)) = 2$.

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

$$\sum_{i=l}^{j} a_i = 0, \qquad \text{if } j < l$$

and

$$\binom{t}{i} = 0, \qquad \text{if } i < 0 \text{ or } i > t. \tag{9}$$

A. Substitutions

In the case when $V = A_q^n$ 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 A_q^n (see Example 2.1). Set

$$N(A_q^n, t) = N_H(A_q^n, t)$$

in this case. The problem is to find $N(A_q^n, t)$ and a simple algorithm for the reconstruction of $x \in A_q^n$ with the help of any different $y_1, \ldots, y_N \in B_t(x)$ where $N = N(A_q^n, t) + 1$.

Theorem 1: For any n, q, and t

$$N(A_q^n, t) = q \sum_{i=0}^{t-1} \binom{n-1}{i} (q-1)^i. \tag{10}$$

If $y_j = y_{1,j} \cdots y_{n,j}$, $j = 1, \ldots, N = N(A_q^n, t) + 1$, are different words in $B_t(x)$ for some $x = x_1 \cdots x_n \in A_q^n$, then

$$x_i = m_q(y_{i,1}, \ldots, y_{i,N}), \qquad i = 1, \ldots, n. \tag{11}$$

Proof: Let all letters of the words $x, z \in A_q^n$ 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 $N(A_q^n, t)$ is not smaller than the right-hand side of (10). The opposite inequality will follow from the proof of the second part of the theorem. To this end, we note that for any $i = 1, \ldots, n$ and $a \in A_q$, $a \neq x_i$, the number of $v = v_1 \cdots v_n \in B_t(x)$ such that $v_i = a$ equals $\sum_{j=0}^{t-1} \binom{n-1}{j} (q-1)^j$. Therefore, the letter x_i occurs more often than others among $y_{i,1} \cdots y_{i,N}$ and we can apply the majority function to find x_i. □

Thus, Theorem 1 determines the minimum number $N = N(A_q^n, t) + 1$ such that N different words obtained from a word $x \in A_q^n$ by at most t substitutions are always sufficient for its reconstruction. Moreover, 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 numerical example has been given in Section I.

B. Transpositions and Asymmetric Errors

In the case when $V \subseteq A_q^n$ 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 $x \in A_q^n$, i.e., $k(z) = k(x)$, see Example 2.4 and (3). In particular, for $V = A_2^n$ each word in $B_t(x)$ has the same Hamming weight. Therefore, it is natural to consider the problem to find $N^\circ(J_w^n, t)$ where J_w^n consists of all words $x \in A_2^n$ with w ones and $n-w$ zeros. If we define the distance between elements of J_w^n to be 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

$$N^\circ(J_w^n, t) = n \sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i} \frac{1}{i+1}. \tag{12}$$

If $y_j = y_{1,j} \cdots y_{n,j}$, $j = 1, \ldots, N = N^\circ(J_w^n, t) + 1$, are different words in $B_t(x)$ for some $x = x_1 \cdots x_n \in J_w^n$, and $u_i = y_{i,1} \cdots y_{i,N}$, $i = 1, \ldots, n$, then

$$x_i = f_{\tau_0, \tau_1}(k(u_i)), \qquad i = 1, \ldots, n \tag{13}$$

where

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

Proof: Denote the right-hand side of (12) 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

$$\begin{aligned} &|B_t(x) \cap B_t(z)| \\ &\quad = 2\sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w-1}{i} + \sum_{i=0}^{t-1} \binom{w-1}{i} \\ &\qquad \cdot \binom{n-w-1}{i+1} + \sum_{i=0}^{t-1} \binom{w-1}{i+1} \binom{n-w-1}{i} \\ &\quad = \sum_{i=0}^{t-1} \binom{w-1}{i} \binom{n-w}{i+1} + \sum_{i=0}^{t-1} \binom{w}{i+1} \\ &\qquad \cdot \binom{n-w-1}{i} = N^\circ. \end{aligned}$$

This proves that $N^\circ(J_w^n, t) \geq N^\circ$. The opposite inequality will follow from the proof of the second part of the theorem. To prove this part, we note that for any $x = x_1 \cdots x_n \in J_w^n$ and i, $i = 1, \ldots, n$, the number of $v = v_1 \cdots v_n \in B_t(x)$ such that $v_i \neq x_i$ 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 (13) 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 $N = N^\circ(J_w^n, t) + 1$ such that N different words obtained from a word $x \in J_w^n$ by at most t transpositions are always sufficient for its reconstruction. Moreover, 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 numerical example has been given in Section I .

It is worth pointing out that any $n+1$ words

$$y_j = y_{1,j} \cdots y_{n,j}, \qquad j = 1, \ldots, n+1$$

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

$$x_i = f_{\tau_0, \tau_1, \ldots, \tau_{q-1}}(k(u_i)), \qquad i = 1, \ldots, n \tag{14}$$

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

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

It follows from the fact that the number of $v = v_1 \cdots v_n \in B_1(x)$ such that $v_i = a \in A_q^n$ 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 $x \in A_3^7$ by at most one transposition of two symbols. Any column allows us to determine the composition of $x : k(x) = (2, 3, 2)$. Using (14) we find that $x = 2011201$.

Consider asymmetric errors

$$H = \{h_i^{>}: i = 1, 2, \ldots\}$$

and

$$H = \{h_i^{<}: i = 1, 2, \ldots\}$$

(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 $y_1, \ldots, y_N$ belong to $B_t(x, H)$ 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

$$y_j = y_{1,j} y_{2,j} \cdots y_{n,j}, \qquad j = 1, \ldots, N$$

then

$$x_i = y_{i,1} \vee y_{i,2} \vee \cdots \vee y_{i,N}, \qquad i = 1, \ldots, n$$

or, respectively,

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

The proof of Theorem 3 is a simple modification 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 $V = A_q^n$ 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 $x \in A_q^n$ in the deletion/insertion metric introduced in [12] (see also [14]). Since the length of words of $B_t(x)$ varies from $n-t$ to $n+t$, we consider separately the case of exactly t deletions and the case of exactly t insertions. Let

$$D_t(x) = B_t(x) \cap A_q^{n-t} \qquad I_t(x) = B_t(x) \cap A_q^{n+t}. \tag{15}$$

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

$$N_q^{-}(n, t) = \max_{x, z \in A_q^n;\, x \neq z} |D_t(x) \cap D_t(z)| \tag{16}$$

and

$$N_q^{+}(n, t) = \max_{x, z \in A_q^n;\, x \neq z} |I_t(x) \cap I_t(z)|. \tag{17}$$

Moreover, we shall describe simple algorithms that recover $x \in A_q^n$ 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 detailed proof of these results is omitted here, it will be published in the *Journal of Combinatorial Theory*.

For any nonnegative integers n and t define

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

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 $A_q = \{0\}$. In this case, $D_q(n, t) = 1$ if $n \geq t \geq 0$. As was mentioned in [4] (with the reference to the report [3]) Calabi proved that

$$D_q(n, t) = \sum_{i=0}^{q-1} D_q(n-i-1, t-i), \qquad \text{when } n > t \geq 0. \tag{18}$$

Recently, Hirschberg [10] found a recurrence on q, $q \geq 2$, for computing $D_q(n, t)$, namely

$$D_q(n, t) = \sum_{i=0}^{t} \binom{n-t}{i} D_{q-1}(t, t-i). \tag{19}$$

In particular

$$\begin{aligned} D_2(n,\,t) &= \sum_{i=0}^{t} \binom{n-t}{i}, \\ D_3(n,\,t) &= \sum_{i=0}^{t} \binom{n-t}{i} \sum_{j=0}^{t-i} \binom{i}{j}. \end{aligned} \tag{20}$$

The formula for $q=2$ is due to Calabi [3].

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

$$N_q^-(n,\,t) = \sum_{i=1}^{q-1} D_q(n-i-1,\,t-i) + D_q(n-2,\,t-1). \tag{21}$$

Using (18) we have that for $n \geq t+1$

$$N_q^-(n,\,t) = D_q(n,\,t) - D_q(n-1,\,t) + D_q(n-2,\,t-1) \tag{22}$$

and for $n \geq t+2$

$$N_q^-(n,\,t) = \sum_{i=0}^{q-1} N_q^-(n-i-1,\,t-i). \tag{23}$$

From (19) and (22) it follows that

$$N_q^-(n,\,t) = \sum_{i=0}^{t-1} \binom{n-t-1}{t-i-1} \left(D_{q-1}(t,\,i) + D_{q-1}(t-1,\,i)\right) \tag{24}$$

and one can use (19) and (20) to calculate $N_q^-(n,\,t)$. In particular

$$N_2^-(n,\,t) = 2\sum_{i=0}^{t-1} \binom{n-t-1}{i} \tag{25}$$

$$\begin{aligned} N_3^-(n,\,t) &= \sum_{i=0}^{t-1} \binom{n-t-1}{i} \sum_{j=0}^{t-i-1} \left(\binom{i+1}{j} + \binom{i}{j} \right) \\ N_q^-(n,\,0) &= 0 \qquad N_q^-(n,\,1) = 2 \\ N_q^-(n,\,2) &= 2n-3-\delta_{q,\,2} \\ N_q^-(n,\,3) &= (n-2)^2 - (3n-10)\delta_{q,\,2} - \delta_{q,\,3} \end{aligned} \tag{26}$$

where $\delta_{i,\,j}$ is the Kronecker symbol. Since

$$D_{q-1}(t,\,i) \leq (q-1)^{t-i}$$

(24) also implies (1).

Now we describe an *algorithm for reconstruction* of an unknown $x = x_1 \cdots x_n \in A_q^n$ when we know a set

$$\{y_1,\,\ldots,\,y_N\} \subseteq D_t(x)$$

(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

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

and $N_q^-(t+1,\,t) \geq |D_t(x)|$ for any $x \in A_q^{t+1}$. 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 a smaller number of rows.

We shall consider the words

$$y_j = y_{1,\,j}y_{2,\,j},\,\ldots,\,y_{n-t,\,j}, \qquad j = 1,\,\ldots,\,N$$

as columns of a matrix Y of size $(n-t)\times N$. For any $a \in A_q$, 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

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

and the ordered composition

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

of u (see (5) and (4)). Consider the thresholds

$$\tau_i = N_q^-(n-i-1,\,t-i), \qquad i = 0,\,1,\,\ldots,\,q-1 \tag{27}$$

and note that due to (23) the threshold function $f_{\tau_0,\,\tau_1,\,\ldots,\,\tau_{q-1}}$ is defined on $l(u)$. If $f_{\tau_0,\,\tau_1,\,\ldots,\,\tau_{q-1}}(l(u)) = j$, then $0 \leq j \leq t$ and

$$x_i = \theta_{i-1}, \qquad \text{for } i = 1,\,\ldots,\,j+1. \tag{28}$$

In the case $j = t$ the submatrix Y'_{θ_j} consists of the unique column $x_{j+2}\cdots x_n$, and the reconstruction of x is completed. In the case $j < t$ 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 $n-j-1$ from $N_q^-(n-j-1,\,t-j)+1$ of 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 (26), and 25 columns of the matrix at the bottom of this page be subsequences of a certain $x = x_1\cdots x_7 \in A_3^7$. For the first row $u \in A_3^{25}$ 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 $f_{\tau_0,\,\tau_1,\,\tau_2}(l(u)) = 1$. Hence $x_1 = \theta_0 = 2$, $x_2 = \theta_1 = 0$, and the problem is reduced to reconstruction $x_3\cdots x_7 \in A_3^5$

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

from $N_3^-(5,\,2)+1=8$ 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 $u \in A_3^8$ 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 $f_{\tau_0,\,\tau_1,\,\tau_2}(l(u)) = 2$. Hence $x_3 = \theta_0 = 1$, $x_4 = \theta_1 = 2$, $x_5 = \theta_2 = 0$. 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 $x \in A_q^n$ and

$$|I_t(x)| = I_q(n,\,t), \qquad \text{for any } x \in A_q^n \tag{29}$$

where

$$I_q(n,\,t) = \sum_{i=0}^{t} \binom{n+t}{i} (q-1)^i. \tag{30}$$

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 $n \ge 0$ and $t<0$, and that agrees with (30) (see (9)). Since for any $x = x_1 \cdots x_n \in A_q^n$

$$I_t(x) = x_1 I_t(x_2 \cdots x_n) \cup \bigcup_{a \in A_q \setminus \{x_1\}} a I_{t-1}(x_1 \cdots x_n) \tag{31}$$

(here $aU = \{z = au\colon\ u \in U\}$ for any $a \in A_q$ and any $U \subseteq A_q^{n+t-1}$), this fact can be proved by induction on $n+t$ ($n \ge 1$, $t \ge 1$), and (30) can be easily found by calculation of $|I_t(0^n)|$. Note that from (31) it also follows that for $n \ge 1$

$$I_q(n,\,t) = I_q(n-1,\,t) + (q-1) I_q(n,\,t-1) \tag{32}$$

and hence

$$I_q(n,\,t) = \sum_{i=0}^{t} I_q(n-1,\,t-i)(q-1)^i \tag{33}$$

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

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

$$N_q^+(n,\,t) = \sum_{i=0}^{t-1} \binom{n+t}{i} (q-1)^i (1-(-1)^{t-i}). \tag{34}$$

Using

$$\binom{n+t}{i} = \binom{n+t-1}{i} + \binom{n+t-1}{i-1}$$

we see that for $n \ge 1$

$$N_q^+(n,\,t) = N_q^+(n-1,\,t) + (q-1) N_q^+(n,\,t-1) \tag{35}$$

and hence

$$N_q^+(n,\,t) = \sum_{i=0}^{t} N_q^+(n-1,\,t-i)(q-1)^i \tag{36}$$

since $N_q^+(n,\,0)=0$. Note that for $n \ge 1$, $t \ge 0$

$$I_q(n,\,t) - N_q^+(n,\,t) = \sum_{i=0}^{t} \binom{n-1+i}{i} (q-1)^i (q-2)^{t-i}$$

which can be proved by induction on $n+t$ with the help of (32) and (35). It follows that

$$|I_t(x)| > N_q^+(n,\,t), \qquad \text{for any } x \in A_q^n,\ n \ge 1,\ t \ge 0. \tag{37}$$

Now we describe an *algorithm for reconstruction* of an unknown $x = x_1 \cdots x_n \in A_q^n$ with the help of any set $\{y_1,\,\ldots,\,y_N\}$ from $N = N_q^+(n,\,t)+1$ different elements of $I_t(x)$ (i. e., its supersequences of length $n+t$), $n \ge 1$, $t \ge 1$. We consider the words

$$y_j = y_{1,\,j} y_{2,\,j} \cdots y_{n+t,\,j}, \qquad j = 1,\,\ldots,\,N$$

as columns of a matrix Y of size $(n+t) \times N$. For any $a \in A_q$ and $i = 1,\,2,\,\ldots,\,n+t$, 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

$$|Y'_{a,\,i}| \ge |Y_{a,\,i}| (q-1)^{-i+1}.$$

For any $b \in A_q$ consider the vector $w(b) \in R^{t+1}$ defined as

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

It is clear that

$$\sum_{i=0}^{t} w_i(b) = N_q^+(n,\,t) + 1 \tag{38}$$

for $b = x_1$. In the case when there exists $b \in A_q$, $b \ne x_1$, for which (38) also holds (and hence this b occurs among the first $t+1$ positions of any column of Y), one can show that the number of columns in which the first occurrence of x_1 precedes that of b is larger than $|Y|/2$. This allows us to find x_1 with the help of the first $t+1$ rows of the matrix Y. Consider the thresholds

$$\tau_i = N_q^+(n-1,\,t-i)(q-1)^i, \qquad i = 0,\,1,\,\ldots,\,t \tag{39}$$

and note that due to (36) and (38) with $b = x_1$ the threshold function $f_{\tau_0,\,\tau_1,\,\ldots,\,\tau_t}$ is defined on $w(x_1) \in R^{t+1}$. If

$$f_{\tau_0,\,\tau_1,\,\ldots,\,\tau_t}(w(x_1)) = j$$

then $0 \le j \le t$ 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 $I_{t-j}(x_2 \cdots x_n)$. In the case $j=t$, $Y'_{x_1,\,j+1}$ consists of the only column $x_2 \cdots x_n$ and this completes the reconstruction of x. In the case $0 \le j < t$, we determine x_1 and reduce the problem considered to the 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 the first $t+1$ rows of the matrix Y. In general, this letter differs from the letter obtained by applying the majority function to the first row of Y. The number $j = f_{\tau_0, \tau_1, \ldots, \tau_t}(w(x_1))$, $0 \le j \le t$, 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 that (34) implies 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 (see the bottom of this page) derived from 21 supersequences of an unknown $x = x_1x_2x_3 \in A_3^3$. We have $w(0) = (9, 2, 5)$, $w(1) = (9, 2, 1)$, $w(2) = (3, 14, 4)$, and see that (38) holds only for $b = 2$, and hence $x_1 = 2$. Since $\tau_0 = 16$, $\tau_1 = 4$, $\tau_2 = 0$, we get $f_{16, 4, 0}(w(2)) = 1$. The matrix $Y_{2,2}$ has 14 columns and $Y'_{2,2}$ has seven different columns each of which is obtained from x_2x_3 by a single insertion. However, for further reconstruction, it is sufficient to use any $N_3^+(2, 1)+1 = 3$ of them, for instance

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

Now we have $w(0) = (2, 1)$, $w(1) = (1, 2)$, $w(2) = (0, 0)$, and there exist two letters 0 and 1 for which (38) hold. However, the first occurrence of 0 precedes that of 1 in a larger number of columns and hence $x_2 = 0$. Then we conclude that $\tau_0 = 2$, $\tau_1 = 0$, $f_{2,0}(w(0)) = 1$, and $Y'_{0,2}$ consists of 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. Given number t of possible errors, the problem of reconstruction of arbitrary vertices within distance ρ is close to an exact reconstruction of vertices belonging to a subset of vertices (code) with the minimum distance $2\rho+1$ between its different elements. This gives rise to new problems for ρ-error-correcting codes when the number t of possible errors exceeds ρ. Although these problems make 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 graph with a finite set V of vertices and a set E of edges which are unordered pairs of distinct elements of V. We denote by $\rho(x, y)$ the *path metric* of Γ, 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, $i = 0, 1, \ldots, s$, we consider the *metrical spheres of radius* i

$$S_i(x) = \{y \colon y \in V, \rho(x, y) = i\} \tag{40}$$

(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

$$p_{i,j}(x, y) = |S_i(x) \cap S_j(y)|. \tag{41}$$

Connected graphs Γ for which values (41) depend only on i, j, and $\rho(x, y)$ are called *distance-regular.* For such graphs, we denote $p_{i,j}(x, y)$ by $p_{i,j}^k$ if $\rho(x, y) = k$. There exists a 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.

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

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

$$I_k(V;\, t) = \max_{x,\, y \in V, \rho(x,\, y)=k} |B_t(x) \cap B_t(y)| \tag{42}$$

and

$$N(V;\, t,\, d) = \max_{d \le k \le s} I_k(V;\, t). \tag{43}$$

The last 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 the other. In particular, $N(V;\, t,\, d) = 0$ for $d \ge 2t + 1$.

The number $N(V;\, t,\, d)$ will be significant for the following problem: Given a code $C \subseteq V$ (any subset of V), and $0 \le \rho \le t$, what is the minimum integer $M = M(C;\, t,\, \rho)$ such that for each $x \in C$, any M vertices in $B_t(x)$ ("erroneous patterns of x") suffice to recover x within distance ρ (or exactly if $\rho = 0$)? Formally, we define $M(C;\, t,\, \rho)$ as the minimum integer M such that for any set $Y \subseteq V$ of size M there exists $z \in V$ satisfying $\rho(x,\, z) \le \rho$ for every $x \in C$ with $Y \subseteq B_t(x)$ (if such x exists). Such a number M always exists under the assumption that $B_t(x)$ and $B_t(y)$ are different for different $x,\, y \in V$. We also assume that C is not contained in $B_\rho(z)$ for any $z \in V$. Under these 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 the case $\rho = t$ any element of $Y \subseteq B_t(x)$ can be chosen as permissible approximation.

First we consider the main case $C = V$.

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

$$M(V;\, t,\, \rho) \ge N(V;\, t,\, 2\rho + 1) + 1 \tag{44}$$

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

Proof: The case $\rho = t$ has been considered. By the definition of $N(V;\, t,\, 2\rho + 1)$, there exist $x,\, y \in V$ such that $\rho(x,\, y) \ge 2\rho + 1$ and $|B_t(x) \cap B_t(y)| = N(V;\, t,\, 2\rho + 1)$. For $Y = B_t(x) \cap B_t(y)$ this contradicts the existence of $z \in V$, for which $\rho(x,\, z) \le \rho$ and $\rho(y,\, z) \le \rho$, and gives (44). On the other hand, if $|Y| \ge N(V;\, t,\, 2\rho + 1) + 1$ for some $Y \subseteq V$, then the distances between different $x \in C$ with $Y \subseteq B_t(x)$ (if they exist) are not greater than 2ρ. Therefore, for $\rho = 0$ there exists at most one such x, and we have equality in (44). □

Thus

$$M(V;\, t,\, 0) = N(V;\, t,\, 1) + 1 \tag{45}$$

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 $W \subseteq V$ 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 $x \in W$. Nevertheless, one may expect that for some graphs, in particular, for the Hamming and Johnson graphs, equality in (44) takes place for all ρ.

The value $N(V;\, t,\, d) + 1$ is also relevant for exact reconstruction of vertices in a code $C \subseteq V$ of minimum distance

$$d(C) = \min_{x,\, y \in C, x \ne y} \rho(x,\, y). \tag{46}$$

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

$$M(C;\, t,\, 0) \le N(V;\, t,\, d) + 1 \tag{47}$$

with equality for some codes (for a sufficient condition of equality see Lemma 3).

Proof: In the 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 definitions of $N(V;\, t,\, d)$ and $d(C)$, for any $Y \subseteq V$ such that $|Y| = N(V; t, d) + 1$ there exists at most one vertex $x \in C$ with $Y \subseteq B_t(x)$. This implies (47). On the other hand, there exist $x,\, y \in V$ such that $|B_t(x) \cap B_t(y)| = N(V;\, t,\, d)$ and $\rho(x,\, y) \ge d$. Therefore, if these x and y belong to a code C, then one cannot exactly reconstruct a vertex of C when knowing $Y = B_t(x) \cap B_t(y)$, and hence $M(C; t,\, 0) > N(V;\, t,\, d)$. □

For calculation of $N(V;\, t,\, d)$ we use a property of graphs which reflects our intuitive expectation that $I_k(V;\, t)$ must decrease with increasing k. A graph Γ is called *monotone on intersections* if for any t, $t = 1,\, \ldots,\, s$, the value $I_k(V;\, t)$ does not increase with k, $k = 1,\, \ldots,\, s$. If Γ has this property, then

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

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

Using the 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

$$N(V;\, t,\, d) = \sum_{i=0}^{t} \sum_{j=0}^{t} p_{i,\, j}^{d} \tag{48}$$

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 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*. This means that for any single error $h \in H$ and $x,\, y \in V$, $x \ne y$, we have $h(x) \ne h(y)$ if h is defined on x and y. 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 will write this property as $H(V) = H^{-1}(V)$. Note that $H(V) = H^{-1}(V)$ holds if $H = H^{-1}$, i.e., $h \in H$ if and only if $h^{-1} \in H$. However, we will verify that this is not a necessary condition for $H(V) = H^{-1}(V)$ to hold. A single error $h \in H$ is called an *involution* if $h^{-1} = h$. In particular, $H = H^{-1}$ takes place if H consists of involutions. If a single error $h \in H$ is defined on all $x \in V$, then h is simply a *permutation* of V. Let us construct a graph $\Gamma_H = \Gamma(V, E)$ 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 in what follows, hx stands for $h(x)$. The property $H(V) = H^{-1}(V)$ implies the crucial fact that the path metric $\rho(x, y)$ on 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 type H).

Note that the definition of Γ_H does not depend on whether $hx = x$ or h is not defined on x. Therefore, one can assume without loss of generality that for any $h \in H$ there exists $x \in V$ such that $hx \neq x$. It is in general not true that if we put $hx = x$ for each $x \in V$ which $h \in H$ is not defined on, we obtain a permutation of V. On the other hand, any single error $h \in H$, which is a permutation of V, can be given by a product of cycles of length 2 or more (with omitted cycles of length one). In particular, the Petersen graph in Fig. 2 is a graph Γ_H with the set H of four single errors (involutions):

$$(01)(23)(57)(69),\ (12)(34)(68)(79),\ (04)(16)(27)(58),\ (05)(38)(49)$$

or with the set H of three single errors:

$$(01234)(57968),\ (04321)(58697),\ (05)(16)(27)(38)(49).$$

In both cases we have $H = H^{-1}$.

The construction of the graphs $\Gamma_H = \Gamma(V, E)$ with $V = A_q^n$ can be used for many types $H = H_n$ of single errors considered in coding theory (see also [12], [13], [18]). In the case $H(V) = H^{-1}(V)$, 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 obtain 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 Sections 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 A_q^* (Examples 2.5 and 2.6) the property $H = H^{-1}$ is also satisfied, and the graph Γ_H generates the deletion/insertion metric on A_q^* [12]. Any $x \in A_q^*$ belongs to the range of definition of a finite number of single errors $h \in H$ and hence all metric spheres (40) are finite for this countable graph. In this case, we have found

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

The restriction $H(V) = H^{-1}(V)$ 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 we 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 A_q^* (Example 2.2) it is not true that $H(V) = H^{-1}(V)$. However, in both cases the parallelogram property holds, and we get the same graph Γ_H on the set A_q^*. In Section II-B for this graph we have, 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.

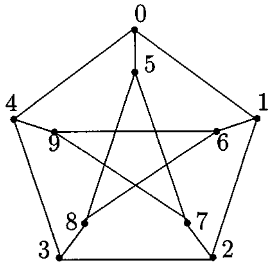

Fig. 2. The Petersen graph.

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, $H = H^{-1}$ and all single errors of H are permutations of V. They all are automorphisms of Γ_H if V is an Abelian group. A graph $\Gamma_H = \Gamma(V, E)$ with Abelian group V and $H \subseteq V$ is referred to as an *Abelian graph.* Many types of errors, such as substitutions, bursts, cyclic, and arithmetic errors, give rise to Abelian graphs Γ_H.

It is worth pointing out that any graph Γ is a graph Γ_H with error metric. For instance, we can consider each edge of Γ as an involution which is defined only on two vertices and permutes them. It is, therefore, of interest to minimize the size of a set H of single errors for which a graph Γ coincides with Γ_H. It is clear that $|H| \geq r$ for any graph Γ of degree r. In the sequel, we consider finite graphs Γ.

Lemma 4: Any graph Γ of degree r is a graph Γ_H for a set H consisting of $r+1$ involutions. There exist graphs Γ of degree r which cannot be represented as graphs Γ_H for a set H consisting of r involutions.

Proof: A graph $\Gamma = \Gamma(V, E)$ coincides with a graph Γ_H where H consists of m involutions if and only if there exists a partition of Γ into m subgraphs $\Gamma(V, E_i)$, $i = 1, \ldots, m$, of degree one. This reduces the problem under consideration to the known problem of coloring edges of a graph using the minimum number m of colors letting the colors of any adjacent edges be different. This problem for a graph of degree r with parallel edges was solved by Shannon in [21]. He proved that $m \leq \lfloor 3r/2 \rfloor$ and showed that this bound is tight for any $r \geq 2$. For the class of graphs without parallel edges that we are interested in, the problem was solved by Vizing in [23]. He proved that

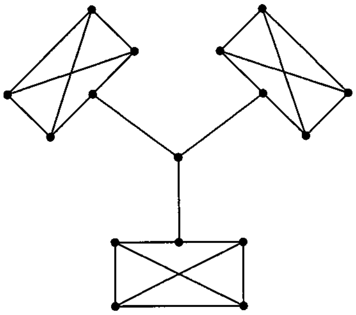

Fig. 3. A regular graph of degree 3 without 1-factors

$m \le r + 1$ and, for any $r \ge 2$, constructed graphs of degree r for which this bound is attained. □

The graphs in Figs. 2 and 3 are examples of regular graphs of degree 3 which cannot be represented as graphs Γ_H with three involutions. However, we have verified that the Peterson graph is a graph Γ_H with a set H of three single errors such that $H = H^{-1}$. Is it true that any graph Γ of degree r is a graph Γ_H for a set H of r permutations such that $H = H^{-1}$? To obtain an answer to this question, we present some facts. Let a regular graph $\Gamma = \Gamma(V, E)$ of degree r be a graph Γ_H with the set H of r single errors such that $H = H^{-1}$. For any $h \in H$ consider the set $E_h = \{\{x, y\}\colon hx = y\}$ and note that $E_h = E_{h^{-1}}$. In the case $h \ne h^{-1}$, $\Gamma(V, E_h)$ is a regular subgraph of Γ of degree two (or 2-factor) if and only if h is a permutation of V without unit cycles and cycles of length two. In the case $h = h^{-1}$, $\Gamma(V, E_h)$ is a regular subgraph of Γ of degree one (or 1-factor) if and only if h is a permutation of V without unit cycles. On the other hand, by the Petersen theorem (see, for instance, [9]), any regular graph of even degree r is partitioned into $r/2$ 2-factors. Since any graph can be converted to a regular graph of the same degree by adding vertices and edges, these arguments show that *any graph* $\Gamma = \Gamma(V, E)$ of degree r can be represented as a graph Γ_H where $|H| = r$ and $H = H^{-1}$ if and only if r is even or r is odd and Γ has a subgraph $\Gamma(V, E_0)$ of degree one such that $\Gamma(V, E\backslash E_0)$ has degree $r - 1$. In particular, the regular graph of degree 3 in Fig. 3 cannot be represented as a graph Γ_H where $|H| = 3$ and $H = H^{-1}$ since it does not have 1-factors. However, the following statement implies that it is still a graph Γ_H with three single errors (for which $H \ne H^{-1}$).

Lemma 5: Any graph Γ of degree r is a graph Γ_H for a set H consisting of r single errors and, in particular, of r permutations (without unit cycles) of vertices in the case of regular Γ.

Proof: By the above-mentioned facts, it is sufficient to prove this statement for a regular graph $\Gamma = \Gamma(V, E)$ of odd degree r only. Denote by $\overline{\Gamma} = \overline{\Gamma}(V, \overline{E})$ the directed graph for which $(x, y) \in \overline{E}$ if and only if $\{x, y\} \in E$. A directed graph $\overline{\Gamma}(V, \overline{E_0})$ is called a *directed* 1*-factor* (on V) if for any $x \in V$ there exists a unique $y \in V, y \ne x$, such that $(x, y) \in \overline{E_0}$ and a unique $z \in V, z \ne x$, such that $(z, x) \in \overline{E_0}$. (Any directed 1-factor consists of disjoint directed cycles of length two or more which cover V.) For any permutation g of V, consider the directed graph $\overline{\Gamma}(V, \overline{E_g})$ where $(x, y) \in \overline{E_g}$ if and only if $y = gx$ and note that $\overline{\Gamma}(V, \overline{E_g})$ is a directed 1-factor if and only if g is a permutation of V without unit cycles. From the definition of single errors and the property $H(V) = H^{-1}(V)$ it follows that to prove the lemma it suffices to show that, for any regular graph $\Gamma(V, E)$ of degree r, the directed graph $\overline{\Gamma}(V, \overline{E})$ is partitioned into r directed 1-factors. The following elegant proof of this fact, which is valid both for even and odd r, was proposed by A. Brouwer in a discussion of the problem. The *bipartite double* of $\Gamma = \Gamma(V, E)$ (see [2]) is the graph $\Gamma' = \Gamma'(V', E')$ where

$$V' = \{x^{(a)}\colon x \in V, a \in \{0, 1\}\}$$

and $\{x^{(a)}, y^{(b)}\} \in E'$ if and only if $\{x, y\} \in E$ and $a \ne b$. 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) $\Gamma' = \Gamma'(V', E_i')$, $i = 1, \ldots, r$ (see, for instance, [9], [16]). Define directed graphs $\overline{\Gamma}(V, \overline{E_i})$, $i = 1, \ldots, r$, as follows: $(x, y) \in \overline{E_i}$ if and only if $\{x^{(0)}, y^{(1)}\} \in E_i'$. The construction used implies that $\overline{\Gamma}(V, \overline{E_i})$, $i = 1, \ldots, r$, form a partition of $\overline{\Gamma}(V, \overline{E})$ and each $\overline{\Gamma}(V, \overline{E_i})$ is a directed 1-factor on V. □

Thus, any graph Γ can be considered as a graph Γ_H with the minimum number of single errors, which is equal to the degree of Γ.

C. Monotonicity 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 $H(V) = H^{-1}(V)$, 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) $\rho(x, hz) \le \rho(x, z)$ if $\rho(x, z) = \rho(y, z)$.

Proof: Let for given $k \ge 2$ and $t = 1, \ldots, s$ the maximum $I_k(V; t)$ (see (42)) be 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 the lemma, it is sufficient to check that

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

Let there exist $z \in V$ such that

$$z \in B_t(x) \cap B_t(y)$$

but

$$z \notin B_t(x) \cap B_t(hy).$$

It is possible only if $\rho(x, z) \le t$, $\rho(y, z) = t$, 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 $\rho(x, hz) \le t$ by property iii). This means that $hz \in B_t(x) \cap B_t(hy)$ but $hz \notin B_t(x) \cap B_t(y)$ and completes the proof. □

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

$$N(A_q^n;\, t,\, d) = \sum_{i=0}^{t-\lceil \frac{d}{2} \rceil} \binom{n-d}{i} (q-1)^i \times \sum_{k=d-t+i}^{t-i} \sum_{l=d-t+i}^{t-i} \binom{d}{k} \binom{d-k}{l} (q-2)^{d-k-l}.$$

In particular

$$N(A_2^n;\, t,\, d) = \sum_{i=0}^{t-\left\lceil \frac{d}{2} \right\rceil} \binom{n-d}{i} \sum_{k=d-t+i}^{t-i} \binom{d}{k}$$

and

$$N(A_2^n;\, t,\, 2t) = N(A_2^n;\, t,\, 2t-1) = \binom{2t}{t}. \tag{49}$$

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 $x_i = y_i + a \mod q$ and $z_i = y_i$, and the properties i)–iii) hold. Since the Hamming graph is distance-regular, we can use Lemma 3 to find $N(A_q^n;\, t,\, d)$ by calculating $|B_t(x) \cap B_t(y)|$ for any points $x,\, y \in A_q^n$ such that $\rho(x,\, y) = d$. Given $x,\, y,\, z$ in A_q^n, denote by i the number of positions where x and y coincide but differ from z, denote by k the number of positions where x and z coincide but differ from y, and denote by l the number of positions where z and y coincide but differ from x. If $z \in B_t(x) \cap B_t(y)$, then $0 \le i \le n-d$, $0 \le k+l \le d$, $d-k+i \le t$, $d-l+i \le t$, and hence $i \le t - d/2$, $k \le t-i$, $l \le t-i$. This implies Corollary 1 taking account of $\binom{a}{b} = 0$ if $b < 0$ or $b > a$. □

As an example, note that if $C \subseteq A_q^n$ is a code of minimum distance $d = 3$, then the minimum number $M(C;\, t,\, 0) = N(A_q^n;\, t,\, 3) + 1$ of erroneous patterns over the combinatorial channel admitting t substitution errors that always suffice to reconstruct an arbitrary point $x \in C$, for $t = 1$, $t = 2$, and $t = 3$, 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 $C \subseteq A_2^n$ of minimum distance $d = 2t-1$, the equality (49) means that the minimum number of sequences $M(C;\, t,\, 0) = N(A_2^n;\, t,\, 2t-1) + 1$ that are sufficient to reconstruct an arbitrary $x \in C$ for the combinatorial channel with at most t substitutions equals $\binom{2t}{t} + 1$ independent of length n.

Corollary 2: 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 monotonicity 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 $\rho(x,\, hz) \le \rho(x,\, z)$. Then we shall apply Lemma 3 to calculate $N(J_w^n;\, t,\, d)$, since the Johnson graph is distance-regular. Let $I = \{1,\, \ldots,\, n\}$ and $I_x = \{j \in I\colon x_j = 1\}$ for any $x = (x_1,\, \ldots,\, x_n)$. Then

$$|I_x \cap I_y| = w - d$$

and

$$|I_x \cap (I \backslash I_y)| = |I_y \cap (I \backslash I_x)| = d$$

and hence h transposes one element of $I_x \cap (I \backslash I_y)$ with an element of $I_y \cap (I \backslash I_x)$. Since this transposition h applied to z increases its distance from y, it must transpose one element of $I_z \cap I_y \cap (I \backslash I_x)$ with an element of $(I \backslash I_z) \cap (I \backslash I_y) \cap I_x$. However, in this case the action of h on z decreases its distance from x. In order to calculate $|B_t(x) \cap B_t(y)|$ put

$$|I_z \cap I_x \cap (I \backslash I_y)| = a$$
$$|I_z \cap I_y \cap (I \backslash I_x)| = b$$
$$|I_z \cap (I \backslash (I_x \cup I_y))| = i$$

and note that $|I_z \cap I_x \cap I_y| = w - a - b - i$, $d(x,\, z) = i + b$, and $d(y,\, z) = i + a$. This implies the formula above. □

Note that (10) and (12) are special cases of Corollaries 1 and 2 for $d = 1$.

It is worth pointing out that 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 [17]). 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 diameter 7. However, $I_1(V';\, 2) = 44 < I_2(V';\, 2) = 74$ for its vertex set V'.

IV. Probabilistic Channels

The problem of 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 within a certain probability. For channels with continuous 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 with a known probability distribution. However, we shall see that some methods and results of the theory of information transmission can be successfully applied to solve these problems.

A. Optimal N-Recontructors for Discrete Channels

Consider the problem of reconstruction of an unknown sequence $x = (x_1, \dots, x_n) \in A_q^n$ provided that the permissible errors transform x to vectors of A_r^n (of the same length) with some probabilities. The problem is to find the minimum number N such that an arbitrary $x \in A_q^n$ can be reconstructed with a preset accuracy and error probability from N distorted versions of x.

In order to give to this problem a precise formulation, we use the Shannon notion of a discrete memoryless channel with input and output alphabets A_q and A_r, respectively. 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 $j \in A_r$ when the letter $i \in A_q$ is transmitted; $\sum_{j \in A_r} P(j|i) = 1$ for any $i \in A_q$. We shall denote this channel (as its transition matrix) by C. Thus, the probability $P_C(v|x)$ of receiving $v = (v_1, \dots, v_n) \in A_r^n$ when $x = (x_1, \dots, x_n) \in A_q^n$ is transmitted over a discrete memoryless channel C can be expressed as

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

For any $x = (x_1, \dots, x_n) \in A_q^n$ we can consider the sequence $(y_1, \dots, y_N)$ of elements of A_r^n as a sequence of patterns of x distorted by errors in the channel C. Contrary to the case of combinatorial channels, some elements of this sequence might be identical. We again consider

$$y_j = (y_{1,j}, \dots, y_{n,j}) \in A_r^n, \qquad j = 1, \dots, N$$

as columns of a matrix $Y = (y_{i,j})$ over A_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 A_r of the size $n \times N$. Let $\mathcal{F}_N$ be the set of all mappings $F: Y_{n,N} \to A_q^n$, $n = 1, 2, \dots$, which are referred to as *N-reconstructors*. For fixed $x = (x_1, \dots, x_n) \in A_q^n$ we can consider $Y = (y_{i,j})$ as an $Y_{n,N}$-valued random variable with the probability assignment

$$P_C(Y|x) = \prod_{k=1}^{n} \prod_{j=1}^{N} P(y_{k,j}|x_k). \tag{50}$$

For any $0 \le d < n$, and any $F \in \mathcal{F}_N$ one can calculate the error probability

$$P_C(F, x, d, N) = \sum_{Y \in Y_{n,N},\, d_H(F(Y), x) > d} P_C(Y|x) \tag{51}$$

of reconstructing $x \in A_q^n$ within Hamming distance d. Note that the case $d = 0$ corresponds to the exact reconstruction. We set

$$P_C(n, d, N) = \min_{F \in \mathcal{F}_N} \max_{x \in A_q^n} P_C(F, x, d, N) \tag{52}$$

and call an N-reconstructor F *optimal* if it gives the minimum in (52). The function $P_C(n, d, N)$ is a nonincreasing function in N because $\mathcal{F}_N \subset \mathcal{F}_{N+1}$. 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 in some degenerate cases). Thus, $N_C(n, d, \varepsilon)$ is the minimum number of repeated transmissions which allow one to reconstruct any sequence of length n with accuracy up to d letters with 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 on the dependence of ε and d on n.

B. Reducible N-Recontructors for Discrete Channels

For a discrete memoryless channel C it is natural to consider a class of N-reconstructors $(N = 1, 2, \dots)$ whose action reduces N-tuple transmission of a message over C to its single 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 $(n = 1, 2, \dots)$ and $x, z \in A_q^n$

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

We shall use reducible reconstructors to estimate (52).

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

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

Since

$$\sum_{Y \in Y_{n,N},\, F^D(Y) = z} P_C(Y|x) = \prod_{i=1}^{n} P_C(D_{z_i}|x_i) \tag{53}$$

F^D is a reducible N-reconstructor for any channel C and the corresponding transition matrix $C_N = C_{N,D}$ has entries

$$p_{i,j}^{N,D} = P_C(D_j|i), \qquad i, j \in A_q.$$

On the other hand, any reducible N-reconstructor F for a channel C generates a partition $D = (D_0, \dots, D_{q-1})$ of A_r^N (or of some subset thereof) where

$$D_j = \{y: y \in A_r^N, F(y) = j\}.$$

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 (51)] 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 by bounding (52) to restrict our consideration of reducible N-reconstructors to N-reconstructors F^D where D is a partition A_r^N into q subsets.

Denote by $\mathcal{F}_N^\circ$ the set of all reducible N-reconstructors. By analogy with (52) we set

$$P_C^\circ(n, d, N) = \min_{F \in \mathcal{F}_N^\circ} \max_{x \in A_q^n} P_C(F, x, d, N) \tag{54}$$

and call a reducible N-reconstructor F (in particular, $F = F^D$) *optimal* if it gives the minimum in (54) (we shall see that there exists a reducible N-reconstructor which is optimal for all n and d, $0 \le d < n$, whose definition does not depend on d).

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 $i \in A_q$ over this channel $C_{N,D}$ or of its recovery at the output of the N-reconstructor F^D. Let

$$P_C(N) = \min_{D} \max_{i \in A_q} (1 - P_C(D_i|i)) \tag{55}$$

where the minimum is taken over all partitions D of A_r^N into q decoding regions.

Example 4.1: For the channel

$$C = \begin{pmatrix} 1-p & p \\ s & 1-s \end{pmatrix} \tag{56}$$

$N = 2$, and a partition D of A_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 (55) 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 express (54) as a function in $P_C(N)$. For any integers n and d, $0 \le d \le n$, and any real vector $u = (u_1, \ldots, u_n)$ where $0 \le u_i \le 1$, $i = 1, \ldots, n$, consider the function $g(n, d, u) = g(n, d, u_1, \ldots, u_n)$ defined by

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

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

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

Therefore,

$$g(n, d, u_1, \ldots, u_n) \le S\left(n, d, \max_{1 \le i \le n} u_i\right) \tag{57}$$

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

$$P_C^\circ(n, d, N) = S(n, d, P_C(N)). \tag{58}$$

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

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

where $u_i = 1 - P_C(D_{x_i} | x_i)$, $i = 1, \ldots, n$. According to (55) and (57) this implies that

$$P_C(F^D, x, d, N) \le S(n, d, P_C(N)). \tag{59}$$

If one considers the partition D of A_r^N and the letter $i_0 \in A_q$ for which the extrema in (55) are attained, then for $x = i_0^n$ the equality in (59) holds. Using the fact that the minimum in (54) is attained at an N-reconstructor F^D, we get (58). □

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 A_r of the size $n \times N$ into q^n decoding regions $D_z^{(n)}$, $z \in A_q^n$:

$$F(Y) = z \ \text{ if and only if } \ Y \in D_z^{(n)}. \tag{60}$$

Since $P_C(F, x, d, N)$ [see (51)] for $x \in A_q^n$ 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 $\max_{x \in A_q^n} P_C(F, x, d, N)$. 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 A_r^N as follows: $Y = (y_{i,j}) \in D_z^n$ where $z = (z_1, \ldots, z_n)$ if and only if $(y_{i,1}, \ldots, y_{i,N}) \in D_{z_i}$ for every $i = 1, \ldots, n$.

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

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

$$D_{11}^2 = \left\{ \begin{matrix} 11 \\ 11 \end{matrix} \right\}$$

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

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

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

$$P_C^\circ(2, 0, 2) = 1 - \min_{x \in A_2^2} P_C(D_x^2 | x).$$

Since

$$\begin{aligned} &P_C(D_{00}^2 | 00) = (1-p^2)^2 \\ &P_C(D_{01}^2 | 01) = P_C(D_{10}^2 | 10) = (1-p^2)(1-s)^2 \\ &P_C(D_{11}^2 | 11) = (1-s)^4 \end{aligned}$$

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

$$\min_{x \in A_2^2} P_C(D_x^{(2)} | x) > (1-s)^4$$

if, in addition,

$$(1-p^2)^2 - p^2(1-p)^2 > (1-s)^4.$$

All required inequalities hold, 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 $P_C^\circ(n, d, N)$. Still, we shall obtain asymptotically tight bounds to $P_C(n, d, N)$ using reducible reconstructors F^D with partitions D of A_r^N into q subsets D_k (decoding regions of k) satisfying the *maximum-likelihood* (ML) property

$$P_C(y|k) \ge P_C(y|i), \qquad \text{for any } i \in A_q \text{ if } y \in D_k. \tag{61}$$

Note that such a partition D is in general not uniquely defined, however, the value

$$\overline{P}_C(N) = \frac{1}{q}\sum_{i=0}^{q-1}(1 - P_C(D_i|i)) \tag{62}$$

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 A_r^N into q subsets D_i. Therefore (see (55))

$$\overline{P}_C(N) \le P_C(N) \le P_C^*(N) \tag{63}$$

where

$$P_C^*(N) = \min_D \max_{i\in A_q}(1 - P_C(D_i|i)) \tag{64}$$

and the minimum is taken over all partitions D of A_r^N into decoding regions of i, $i \in A_q$, satisfying the ML property.

Example 4.1 (Continued): For the channel (56) with $0 < s < p \le 1/2$ the partition D of A_2^2 above is the unique one satisfying this property. Therefore, using the 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 $s(2 - s) > p$.

The following lemma establishes our basic bounds to $P_C(n, d, N)$. The key to its proof is to show that the average over $x \in A_q^n$ of the probability of error of reconstructing x within distance d is minimized by the reducible N-reconstructor F^D with D satisfying the ML property (for $d = 0$, the latter holds by the standard optimality of ML decoders).

Lemma 8: 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 (58) and (63). To obtain the lower bound we consider for an N-reconstructor F the partition $D^{(n)}$ defined by (60) and note that

$$\max_{x\in A_q^n} \sum_{Y\in Y_{n,N},\, d_H(F(Y),x)>d} P_C(Y|x) \ge \frac{1}{q^n}\sum_{x\in A_q^n}\ \sum_{Y\in Y_{n,N},\, d_H(F(Y),x)>d} P_C(Y|x). \tag{65}$$

With the notation $L_d(z) = \{x \in A_q^n\colon d_H(z, x) > d\}$, we can rewrite (65) as

$$\frac{1}{q^n}\sum_{Y\in Y_{n,N}}\ \sum_{x\in L_d(F(Y))} P_C(Y|x). \tag{66}$$

We shall prove that (66) attains a minimum value when F is the reducible N-reconstructor F^D (or, equivalently, when $D^{(n)} = D^n$), where D is a partition of A_r^N into q regions satisfying the ML property. Then we shall find this minimum value. To this end, for each fixed matrix $Y \in Y_{n,N}$, with rows $u_1, \dots, u_n$, say, we shall show that the minimum of

$$f(z) = \sum_{x\in L_d(z)} P_C(Y|x)$$

is attained for $z = F^D(Y)$. It suffices to check that if $z = z_1\cdots z_n \ne F^D(Y)$, i.e., $u_i \notin D_{z_i}$ for some $0 \le i \le n$ then, replacing z_i with $a \in A_q$ determined by the condition $u_i \in D_a$, z will be changed to $\tilde{z}$ with $f(\tilde{z}) \le f(z)$. To simplify this notation we assume that $i = n$. For $z = z'z_n$, $\tilde{z} = z'a$, where $z' = z_1 \cdots z_{n-1}$, we have $x \in L_d(z)\backslash L_d(\tilde{z})$ if and only if $x = z'a$ for some $x' = x_1 \cdots x_{n-1}$ such that $d_H(x', z') = d$, and $x \in L_d(\tilde{z})\backslash L_d(z)$ if and only if $x = x'z_n$ for some x' as above. By (50), it follows that

$$\begin{aligned} f(z) - f(\tilde{z}) &= \sum_{x\in L_d(z)\backslash L_d(\tilde{z})} P_C(Y|x) - \sum_{x\in L_d(\tilde{z})\backslash L_d(z)} P_C(Y|x) \\ &= \sum_{x':d_H(x',z')=d}\ \prod_{k=1}^{n-1} P_C(u_k|x_k)(P_C(u_n|a) - P_C(u_n|z_n)). \end{aligned}$$

Since the condition $u_n \in D_a$ implies $P_C(u_n|a) \ge P_C(u_n|z_n)$ by the ML property, this completes the proof of the fact that (66) is minimized by $F = F^D$. By (53), $F^D(Y)$ can be regarded as the output of a discrete channel, hence $P_C(F^D, x, d, N)$ represents the probability that, transmitting $x = x_1 \cdots x_n$ over this channel, the output differs from x in more than d components. It follows that (65) for $F = F^D$ represents the probability that the input and output differ in more than d components when a random input, uniformly distributed on A_q^n, is transmitted over this channel. In this case, the input and output differ in the ith component with probability (62), and these events are independent for $i = 1, \dots, n$. This proves that, for $F = F^D$, (65) equals $S(n, d, \overline{P}_C(N))$. □

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 error probability of at most ε for a discrete memoryless channel C.

For any distinct i, $k \in A_q$, consider the set

$$A_{i,k} = \{j \in A_r\colon P(j|i)P(j|k) > 0\} \tag{67}$$

which may be empty. For any s, $0 \le s \le 1$, let

$$\alpha_{i,k}(s) = \sum_{j\in A_{i,k}} (P(j|i))^{1-s}(P(j|k))^s \tag{68}$$

and

$$\alpha(C) = \max_{i,k\in A_q,\, i\ne k}\ \min_{0\le s\le 1} \alpha_{i,k}(s). \tag{69}$$

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 distinct i, $k \in A_q$. This means that any column of 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, $k \in A_q$

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

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 each $j \notin A_{i,k}$. 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 $y \in A_r^n$, $P_C(y|i^n) = P_C(y|k^n)$ where $i^n = (i, \ldots, i) \in A_q^n$, $i \in A_q$. 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 *nondegenerate* channels C whose 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 nondegenerate.

The following statement is derived from the celebrated result on the probability of error for a code with two codewords due to Shannon, Gallager, and Berlekamp [22, Theorem 5]. In fact, we apply their arguments to calculate the corresponding bounds for the "repetition" code $\{i^N : i \in A_q\}$. This explains rather a simple formulation of the statement.

Lemma 9: For any nondegenerate 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 $i, k \in A_q$, $i \ne k$, such that $\alpha(C) = \min_{0 \le s \le 1} \alpha_{i,k}(s)$.

Proof: Consider a partition D of A_r^N into q subsets D_i, $i \in A_q$, satisfying the ML property (61). It follows from (67), (68), and (50) that for each $i, k \in A_q$, $i \ne k$

$$\{y \in A_r^N : P(y|i)P(y|k) > 0\} = A_{i,k}^N$$

and

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

Using that

$$P_C(y|k) \ge P_C(y|i), \qquad \text{for } y \in D_k$$

by the ML property (61), we have for each $i \in A_q$

$$\begin{aligned} 1 - P_C(D_i|i) &= \sum_{k \in A_q \setminus \{i\}} \sum_{y \in D_k} P_C(y|i) \\ &= \sum_{k \in A_q \setminus \{i\}} \sum_{y \in D_k \cap A_{i,k}^N} P_C(y|i) \\ &\le \sum_{k \in A_q \setminus \{i\}} \sum_{y \in A_{i,k}^N} (P_C(y|i))^{1-s_k} (P_C(y|k))^{s_k} \\ &= \sum_{k \in A_q \setminus \{i\}} (\alpha_{i,k}(s_k))^N \end{aligned}$$

for any numbers s_k, $0 \le s_k \le 1$, $k \in A_q \setminus \{i\}$. Choosing s_k so as to minimize $\alpha_{i,k}(s)$, and using the definitions (64) and (69), we obtain the desirable upper bound.

Now fix $i, k \in A_q$, $k \ne i$, such that

$$\alpha(C) = \min_{0 \le s \le 1} \alpha_{i,k}(s)$$

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

$$\alpha_{i,k}(s^\star) = \alpha(C). \tag{70}$$

Since D_i and D_k are disjoint, (62) yields

$$\begin{aligned} q\overline{P}_C(N) &\ge \sum_{y \in A_r^N \setminus D_i} P_C(y|i) + \sum_{y \in A_r^N \setminus D_k} P_C(y|k) \\ &\ge \sum_{y \in A_r^N} \min(P_C(y|i), P_C(y|k)). \end{aligned} \tag{71}$$

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

$$\mu(s) = N \ln \alpha_{i,k}(s) \tag{72}$$

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

$$\mu(s) = \ln \sum_{y \in A_{i,k}^N} (P_C(y|k))^{1-s} (P_C(y|i))^s. \tag{73}$$

The following remarkable fact was established in [22]. If, for any s, $0 \le s \le 1$, one considers the log-likelihood ratio

$$\Delta(y) = \ln \frac{P_C(y|k)}{P_C(y|i)}, \qquad y \in A_{i,k}^N \tag{74}$$

to be a random variable with probability assignment

$$Q_s(y) = \frac{(P_C(y|i))^{1-s} (P_C(y|k))^s}{\sum_{z \in A_{i,k}^N} (P_C(z|i))^{1-s} (P_C(z|k))^s} \tag{75}$$

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

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

we have

$$\sum_{y \in G_s} Q_s(y) \ge \frac{1}{2} \tag{76}$$

by the Chebyshev inequality. Note also that

$$0 \le \mu''(s) = N \frac{\alpha''_{i,k}(s)}{\alpha_{i,k}(s)} - N \left(\frac{\alpha'_{i,k}(s)}{\alpha_{i,k}(s)} \right)^2 \le \frac{N}{2} \beta^2(C). \tag{77}$$

From (73)–(75) it follows that

$$\begin{aligned} P_C(y|i) &= e^{\mu(s)-s\Delta(y)}Q_s(y)\\ P_C(y|k) &= e^{\mu(s)+(1-s)\Delta(y)}Q_s(y) \end{aligned}$$

and hence for any $y \in G_s$

$$\begin{aligned} P_C(y|i) &\geq e^{\mu(s)-s\mu'(s)-s\sqrt{2\mu''(s)}}\, Q_s(y)\\ P_C(y|k) &\geq e^{\mu(s)+(1-s)\mu'(s)-(1-s)\sqrt{2\mu''(s)}}Q_s(y). \end{aligned}$$

According to (71) and (76), to complete the proof it is sufficient to verify that this implies that for any $y \in G_{s^\star}$

$$\min(P_C(y|i),\, P_C(y|k)) \geq (\alpha(C))^N e^{-\beta(C)\sqrt{N}}\, Q_{s^\star}(y). \quad (78)$$

If $\mu'(s)$ changes sign for $0 \leq s \leq 1$, then $\mu'(s^\star) = 0$ and (78) holds due to (70), (72), and (77). If $\mu'(s) \geq 0$ or $\mu'(s) \leq 0$ for $0 \leq s \leq 1$, then $s^\star = 0$ or $s^\star = 1$, respectively, and (78) also holds in both cases. □

Example 4.1 (Continued): For the channel (56) with $0 < s \leq p < 1/2$

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

where

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

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: 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, one can recover any sequence, including 1^n, with arbitrary prescribed error probability, using a sufficient number N of its transmissions over the channel.

Theorem 4: Let $\varepsilon = \varepsilon(n) > 0$ and $d = d(n) \geq 0$ be functions such that $\varepsilon \to 0$ and $d/n \to 0$ as $n \to \infty$. Then for any nondegenerate discrete memoryless channel C

$$N_C(n, d, \varepsilon) \sim \frac{\ln\frac{n}{d+1} + \frac{1}{d+1}\ln\frac{1}{\varepsilon}}{\ln\frac{1}{\alpha(C)}}. \quad (79)$$

Proof: From the definition of the number $N = N_C(n, d, \varepsilon)$ and Lemma 8 it follows that

$$\begin{aligned} S(n, d, \overline{P}_C(N)) &\leq P_C(n, d, N) \leq \varepsilon\\ &< P_C(n, d, N-1) \leq S(n, d, P_C^*(N-1)). \end{aligned} \quad (80)$$

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 the monotonicity of $S(n, d, p)$ on p

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

Therefore, Lemma 9 implies that $N \to \infty$ and

$$-\ln x(n) \sim -N \ln \alpha(C). \quad (81)$$

By Bonferroni's inequality

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

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 (81), this completes the proof. □

By Theorem 4, $N_C(n, d, \varepsilon)$ grows linearly with the 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 bounded when reconstruction of a sequence is permissible with a *fixed fraction* d/n of wrong letters and $\varepsilon = \varepsilon(n)$ is not smaller than an exponent in n. We use the Chernoff bound

$$S(n, d, p) \leq 2^{-T(\delta, p)n}, \qquad \text{if } \delta = \frac{d+1}{n} \geq p \quad (82)$$

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 traverses 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 5: If $d+1 \geq \delta n$, $0 < \delta < 1$, and $\varepsilon \geq 2^{-cn}$, $c > 0$, then for any nondegenerate discrete memoryless channel C and any n, $n = 1, 2, \ldots$

$$N_C(n, d, \varepsilon) \leq \left\lceil \frac{\ln\frac{q-1}{\gamma(\delta, c)}}{\ln\frac{1}{\alpha(C)}} \right\rceil. \quad (83)$$

Proof: Denote by N the right-hand side of (83). Using Lemma 9, we get

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

and hence, by Lemma 8, for $p = P_C^*(N)$, any n and $d+1 \geq \delta n$

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

Note that the Chernoff bound (82) shows that the probability of the event that at least δn wrong letters will result when a sequence of length n is transmitted over a 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 5. In particular, in the case $p = 0.02$, $\delta = 0.01$, and $c = 0.1$ we get that five repetitions are sufficient independently of the length n.

D. Reconstruction for Continuous Channels with Additive Noise

In this subsection 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 that letter and 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 $p(x) = p(-x)$. For any input sequence $\theta = (\theta_1, \ldots, \theta_n) \in R^n$, we can consider sequences $y_1, \ldots, y_N$ in R^n as patterns of θ distorted by errors in the channel C and represent them as columns of a matrix $Y = (y_{i,j})$ over R of the size $n \times N$. Now we denote by $Y_{n,N}$ the set of all real matrices Y of the size $n \times N$ and by $\mathcal{F}_N$ 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 $\theta = (\theta_1, \ldots, \theta_n) \in R^n$ we can consider $Y = (y_{i,j}) \in Y_{n,N}$ as an nN-dimensional random variable with density $\prod_{i=1}^n \prod_{j=1}^N p(x_{i,j} - \theta_i)$. For any $F \in \mathcal{F}_N$ and $\delta > 0$ we can calculate the probability $\Pr(\|F(Y) - \theta\| > \delta)$ of the event that θ is not reconstructed within Euclidean distance δ. (In the sequel, $\|z\|$ stands for the Euclidean norm of $z \in R^n$.) The value

$$P_C(n, \delta, N) = \inf_{F \in \mathcal{F}_N} \sup_{\theta \in R^n} \Pr(\|F(Y) - \theta\| > \delta) \qquad (84)$$

characterizes the maximum error probability of reconstructing θ within Euclidean distance δ for the best (optimal) N-reconstructor. For any n $(n = 1, 2, \ldots)$, $\delta (\delta > 0)$, 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 coincides with the classical statistical problem to find the minimax estimator F of an unknown vector $\theta \in R^n$ based on N independent observations $y_1, \ldots, y_N$ 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

$$w(z) = \begin{cases} 0, & \text{if } 0 \leq z \leq \delta \\ 1, & \text{if } z > \delta \end{cases} \qquad (85)$$

is used. The minimax estimator is defined as the minimizer of

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

where E_θ is the symbol of mathematical expectation for a fixed θ. We mention known results in a general outline (see, for instance, [11], [24]). Under some restrictions on the functions $p(x)$ and $w(z)$, there exists a unique vector $F^* = F^*(Y)$ which minimizes the integral

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

This F^* is called the Pitman estimator and it is minimax in the class of estimators invariant with respect to shifts, i.e., such that $F(y_1 + z, \ldots, y_N + z) = F(y_1, \ldots, y_N) + z$ for all $z \in R^n$. Under some additional restrictions, the Pitman estimator is minimax in the class of all estimators, by the Hunt–Stein theorem. If $p(x)$ is a normal density and $w(z)$ is defined by (85) (or $w(z) = z^2$), then all required conditions hold and

$$F^*(Y) = \frac{1}{N} \sum_{j=1}^N y_j. \qquad (86)$$

Thus, in this case the estimator (86) is minimax for all $n \geq 1$ and $\delta > 0$, although its definition does not depend on δ. (The author has 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

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}}. \qquad (87)$$

In this case, for any n and σ the N-reconstructor (86) is optimal for all $\delta > 0$ and

$$P_G(n, \delta, N) = 1 - \Pr(\|\bar{\xi}\|^2 \leq \delta^2)$$

where

$$\bar{\xi} = \frac{1}{N} \sum_{j=1}^N \xi_j \quad \text{and} \quad \xi_j = (\xi_{1,j}, \ldots, \xi_{n,j}).$$

Each component $(\bar{\xi})_i$, $i = 1, \ldots, n$, of the vector $\bar{\xi}$ is the arithmetical average of N independent and identically distributed (i.i.d.) Gaussian random variables with mean 0 and variance σ^2 and, hence, is Gaussian with mean 0 and variance σ^2/N. Thus, similar to reducible N-reconstructors for a discrete memoryless channel, the action of the optimal N-reconstructor (86) is equivalent to a single transmission of a message $\theta = (\theta_1, \ldots, \theta_n) \in R^n$ over the Gaussian channel with the decreased variance σ^2/N. Moreover, since $(\bar{\xi})_i$, $i = 1, \ldots, n$, are independent random variables, $\sum_{i=1}^n |(\bar{\xi})_i|^2$ has the χ^2 probability distribution

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

and hence

$$P_G(n, \delta, N) = 1 - \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_0^{\frac{\delta^2 N}{2\sigma^2}} \lambda^{\frac{n}{2}-1} e^{-\lambda}\, d\lambda. \qquad (88)$$

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

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

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

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

Proof: For any nonnegative integer m

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

and, hence, (88) implies that for even n

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

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

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

Theorem 6: If σ and δ are fixed, and $\varepsilon = \varepsilon(n) \rightarrow 0$ as $n \rightarrow \infty$, then

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

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

Proof: Because

$$P_G(n,\,\delta,\,N) \leq P_G(n+1,\,\delta,\,N)$$

and hence

$$N_G(n+1,\,\delta,\,\varepsilon) \geq N_G(n,\,\delta,\,\varepsilon)$$

it suffices to consider only even n. Since $Q(\lambda,\,m)$ is a decreasing function of 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 $Q(\lambda,\,\frac{n}{2}-1) = \varepsilon(n)$. Using standard arguments we get for each m

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

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 if $x > 0$ and $\lambda = m + x\sqrt{m}$ with $m \rightarrow \infty$, then $Q(\lambda,\,m)$ is greater than a positive constant if x is a constant, and

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

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

$$-\ln\,Q(\lambda,\,m) \sim \frac{x^2}{2}, \qquad \text{if } x \rightarrow \infty \text{ and } \frac{x}{\sqrt{m}} \rightarrow 0 \tag{89}$$

$$-\ln\,Q(\lambda,\,m) \sim m(\eta - \ln(1+\eta)), \qquad \text{if } \frac{x}{\sqrt{m}} = \eta > 0 \tag{90}$$

$$-\ln\,Q(\lambda,\,m) \sim x\sqrt{m}. \qquad \text{if } \frac{x}{\sqrt{m}} \rightarrow \infty. \tag{91}$$

Let $x(n)$ be defined by $\lambda(n) = m + x(n)\sqrt{m}$ where $m = \frac{n}{2}-1$. Then $x(n) \rightarrow \infty$ since $Q = (\lambda(n),\,m) = \varepsilon(n) \rightarrow 0$. Equations (89)–(91) allow us to find the asymptotic behavior of $x(n)$ and $\lambda(n) = m + x(n)\sqrt{m}$ depending on that of

$$-\frac{1}{n}\ln\varepsilon(n) = -\frac{1}{n}\,\ln\,Q(\lambda(n),\,m).$$

□

In particular, $N_G(n,\,\delta,\,\varepsilon)$ grows linearly with the sequence length n when the permissible error probability ε of reconstruction within a given Euclidean distance δ decreases not faster than exponentially in n.

V. Concluding Remarks and Open Problems

The aim of this paper has been to develop the theory of efficient reconstruction of sequences which deals with optimization problems for 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. We consider repeatedly transmitting an arbitrary message in 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 statistical 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 the 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 $V \subseteq A_q^n$, 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, $V \subseteq A_q^*$) introduced in [12]. At 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 which are permutations of vertices in the case of regular Γ. Coding theory, the theory of sequences, and computational molecular biology (see [1], [6], [8], [19]) give numerous examples of types of single errors (one-to-one partial mappings $V \rightarrow V$) for which the property $H(V) = H^{-1}(V)$ or the weaker parallelogram property is satisfied. An important problem is to describe types of single errors inherent to genes, genomes, and other objects of molecular biology and determine the minimum number of erroneous patterns sufficient for exact reconstruction. It is worth mentioning 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 joining 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 Abelian graphs. An interesting problem is to find the minimum size $M(A_q^n;\,t,\,\rho)$ of a set Y in the Hamming metric ball $B_t(x)$ centered at an arbitrary $x \in A_q^n$ which allows one to approximate this x 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 exact reproduction of 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 the length n of messages, then exact reproduction requires an exponentially increasing number of different erroneous patterns. On the other hand, by Theorem 4, 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 n is sufficient. (The difference is two orders of 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 a proof of the main Theorem 4, and the asymptotic expression (79) is also valid for the class of reducible N-reconstructors. Moreover, this notion gives rise to the new problem of interest to find $P_C(N)$ (see (55)) for a channel C and the corresponding partition of A_r^N into q regions.

The results on the reproduction of a sequence with the help of its repeated transmissions over Gaussian channel lie in the course of traditional problems of mathematical statistics. However, the question of interest is the existence and construction of optimal N-reconstructors whose definition does not depend on δ for the cases when the distribution function of the noise has another symmetric density. Note that Theorems 4–6 allow one to compare asymptotic behavior of the minimum number of repeated transmissions of a message over discrete and continuous memoryless channels that are sufficient to recover this message with a preset accuracy and given error probability.

For probabilistic channels, the problem of reconstructing a sequence when knowing that this sequence belongs to a code $C \subset A_q^n$ is also of interest, but it is not considered in this paper.

Acknowledgment

The author wishes to thank A. Brouwer, I. Csiszár, Y. Golubev, D. Hirschberg, and an anonymous reviewer for comments and suggestions which helped him to improve this paper.

References

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

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

[3] L. Calabi, "On the computation of Levenshtein's distances," Parke Math. Labs., Inc., Carlisle, MA, Tech. Note TN-9-0030, 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. Contr.*, 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, and G. F. Italiano, "Efficient algorithms for sequence analysis," in *Methods in Communication, Security, and Computer Science*. Berlin, Germany: 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, U.K.: Addison-Wesley, 1969.

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

[11] I. A. Ibragimov and R. Z. Hasminskii, *Asymptotic Theory of Estimates* (in Russian). Moscow, U.S.S.R.: Nauka, 1979. English translation: *Statistical Estimation. Asymptotic Theory*. Berlin, Germany: Springer-Verlag, 1981.

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

[13] ——, "Elements of coding theory" (in Russian), in *Discrete Mathematics and Mathematical Problems of Cybernetics*. Moscow, U.S.S.R.: Nauka, 1974, pp. 207–305. German translation in *Diskrete Mathematik und mathematische Fragen der Kybernetik*. Berlin, Germany: Academie-Verlag, 1980, pp. 198–279.

[14] ——, "On perfect codes in deletion/insertion metric" (in Russian), *Discr. Math.*, vol. 3, no. 1, pp. 3–20, 1991. English translation in *Dokl. Math.*, vol. 2, no.3, pp. 241–258, 1992.

[15] ——, "Reconstruction of objects from a minimum number of distorted patterns" (in Russian), *Dokl. Akad. Nauk SSSR*, vol. 354, no. 5, pp. 593–596, 1997. English translation in *Dokl. Math.*, vol. 55, pp.417–420, 1997.

[16] L. Lovász and M. D. Plummer, *Matching Theory*. Budapest, Hungary: Akadémiai Kiadó, 1986.

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

[18] W. W. Peterson and E. J. Weldon Jr., *Error-Correcting Codes*. Cambridge, MA: MIT Press, 1972.

[19] D. Sankoff and J. B. Kruskal, Eds., *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison*. Reading, MA: Addison-Wesley, 1983.

[20] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948.

[21] ——, "A theorem on coloring the lines of a network," *J. Math. Phys.*, vol. 28, pp. 148–151, 1949.

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

[23] V. G. Vizing, "On estimate of the chromatic class of a p-graph" (in Russian), *Discr. Analiz*, vol. 3, pp. 25–30, 1964.

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