V. I. LEVENSHTEIN

1965

SovietRxiv

Abstract

Full Text

V. I. LEVENSHTEIN

BINARY CODES WITH CORRECTION OF DELETIONS, INSERTIONS, AND REVERSALS OF SYMBOLS

(Presented by Academician P. S. Novikov on 4 I 1965)

In studying problems of transmitting binary information through channels, one usually considers a channel model in which failures of the type $0 \to 1, 1 \to 0$, hereafter called reversals, are allowed. In the present paper (as in ([1])) a channel model is studied in which failures of the type $0 \to \Lambda, 1 \to \Lambda$, called deletions, and failures of the type $\Lambda \to 0, \Lambda \to 1$, called insertions, are also allowed (here Λ is the empty word). For such channels, by analogy with the combinatorial problem of constructing optimal codes with correction of s reversals, one considers problems of constructing optimal codes with correction of deletions, insertions, and reversals.

1. Codes with correction of deletions and insertions. Words in the alphabet $\{0, 1\}$ will be called binary words. An arbitrary set of binary words of fixed length will be called a code*. A code K will be called a **code with correction of s deletions** (a **code with correction of s insertions**) if any binary word can be obtained from no more than one word of the code K by means of s or fewer deletions (respectively insertions). A code K will be called a **code with correction of s deletions and insertions** if any binary word can be obtained from no more than one word of the code K by means of s or fewer deletions and insertions. The latter property ensures the possibility of uniquely determining the original code word from the word obtained from it by some number i ($i \geq 0$) of deletions and some number j ($j \geq 0$) of insertions, if $i + j \leq s$. The following assertion shows that all the definitions of codes given above are equivalent.

Lemma 1. *Any code with correction of s deletions (as well as any code with correction of s insertions) is a code with correction of s deletions and insertions.*

Proof (by contradiction). Suppose that from a word x of length n, by means of i_1 deletions and j_1 insertions, where $i_1 + j_1 \leq s$, and from a word y of length n, by means of i_2 deletions and j_2 insertions, where $i_2 + j_2 \leq s$, one obtains the same word z. If in the word z we omit (insert) those symbols which, in obtaining z, are inserted (omitted) in at least one of the words x and y, then, as is easy to see, we obtain a word that can be obtained both from x and from y by no more than $\max(i_2 + j_1, j_2 + i_1)$ deletions (respectively insertions). In

view of the equality of the lengths of the words x and y, $j_1 - i_1 = j_2 - i_2$, and consequently,

$$i_2 + j_1 = j_2 + i_1 = \frac{1}{2}(i_1 + i_2 + j_1 + j_2) \leq s,$$

which proves Lemma 1.

Codes with correction of s deletions and insertions admit another, metric, description. Consider the function $\rho(x, y)$, defined on pairs of binary words and equal to the least number of deletions and insertions that transform the word x into y. It is not difficult to show that the function $\rho(x, y)$ is a metric; moreover, a code K is a code with correction of s deletions and insertions if and only if, for any two distinct words x and y from K, one has $\rho(x, y) > 2s$.

Let B_n be the set of all binary words of length n. For an arbitrary word x from B_n, denote by $|x|$ the number of ones in the word x, and by

---

* The subsequent definitions also make sense if by a code one understands an arbitrary set of words (possibly of different lengths) in some alphabet of r letters ($r \geq 2$). Note, however, that in the case of words of different lengths, Lemma 1 is, generally speaking, no longer true.

$\|x\|$ is the number of runs* of the word x, and let us estimate the number $P_s(x)$ (the number $Q_s(x)$) of distinct words obtained from x by means of s deletions (respectively, s insertions). The following estimates hold:

$$C^s_{\|x\|-s+1} \leq P_s(x) \leq C^s_{\|x\|+s-1}, \tag{1}$$

$$\sum_{i=0}^{s} C^i_n 2^{s-i} \leq Q_s(x) \leq \sum_{i=0}^{s} C^i_n C^i_s 2^{s-i}. \tag{2}$$

To prove the upper estimate in (1), note that each word obtained from x by deletions is uniquely determined by specifying the number of symbols deleted from each run, and, consequently, $P_s(x)$ does not exceed the number of ordered decompositions of the number s into $\|x\|$ nonnegative summands. On the other hand, it is easy to see that if in any s pairwise nonadjacent runs of the word x one deletes one symbol from each, then all words obtained in this way will be distinct. This gives the lower estimate in (1), if one notes that the number of such words is equal to the number of ordered decompositions of the number $\|x\| - s$ into $s + 1$ nonnegative summands, only two of which may be zero. The upper estimate in (2) follows from the fact that each word obtained from the word $x = \sigma_1 \dots \sigma_n$ by s insertions can be obtained in the following way. For some i ($i = 0, 1, \dots, s$), i indices $n_1, \dots, n_i$ are chosen ($1 \leq n_1 < \dots < n_i \leq n$) and $i+1$ words $\beta_1, \dots, \beta_i, \beta_{i+1}$, the sum of whose lengths is equal to s, with each

of the first i words β_j nonempty and not ending with the symbol σ_{n_j}; then each word β_j $(j = 1, \dots, i)$ is inserted into the word x before the symbol σ_{n_j}, and the word β_{i+1} after the symbol σ_n. The lower estimate in (2) follows from the fact that if each of the words $\beta_1, \dots, \beta_i$ has length one, then all the words obtained from x by the method indicated above are distinct.

Note that from (1) and (2) it follows that $P_1(x) = \|x\|$ and $Q_1(x) = n + 2$. Denote by $L_s(n)$ the cardinality (number of words) of a maximal code in B_n with correction of s deletions and insertions.

Lemma 2.** For fixed s and $n \to \infty$,

$$2^s (s!)^2 2^n / n^{2s} \lesssim L_s(n) \lesssim s!\, 2^n / n^s. \tag{3}$$

Proof. Let K be a maximal code in B_n correcting s deletions and insertions, and let, for arbitrary k $(1 \le k < n/2)$,

$$L_s(n) = L'_k + L''_k,$$

where L'_k is the number of words x of the code K such that $k < \|x\| < n - k$. From the definition of the code K it follows that

$$\sum_{x \in K} P_s(x) \le 2^{n-s},$$

and from its maximality,

$$\sum_{x \in K} R_{2s}(x) \ge 2^n,$$

where $R_{2s}(x)$ is the number of words at distance $2s$ or less (in the metric $\rho(x, y)$) from the word x. Using (1) and (2), we obtain

$$2^{n-s} \ge L'_k C^s_{k-s}$$

and

$$2^n \le (L'_k C^s_{n-k+s} + L''_k C^s_{n+s-1}) \sum_{i=0}^{s} C^i_{n-1} C^i_s 2^{s-i}.$$

The estimates (3) follow from the last inequalities, if one notes that

$$L''_k \le 2 \left(\sum_{i=1}^{k} C^{i-1}_{n-1} + \sum_{i=n-k}^{n} C^{i-1}_{n-1} \right) = 2 \sum_{i=0}^{k} C^i_n$$

(since the number of words from B_n having i runs is equal to $2C_{n-1}^{i-1}$), and uses the fact that

$$\sum_{i=0}^{k} C_n^i = o\left(\frac{2^n}{n^{2s}}\right)$$

for $k = [n/2 - \sqrt{sn \ln n}\,]$ and $n \to \infty$ (see, for example, (2)).

---

* A run of the word x is a maximal subword of the word x consisting of identical symbols. For example, the word $x = 01101$ has 4 runs.

** In what follows, the notation $f(n) \lesssim g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) \leq 1$, and the notation $f(n) \sim g(n)$ means that $\lim_{n\to\infty} f(n)/g(n) = 1$.

Theorem 1.

$$L_1(n) \sim 2^n/n. \tag{4}$$

Proof. By Lemma 2, it is enough for us to show that

$$L_1(n) \geqslant 2^n/(n+1). \tag{5}$$

To prove this, we use a construction of Varshamov–Tenengolts ([3]). Consider the class of codes $K_{n,m}^a$, where each $K_{n,m}^a$ $(a = 0, 1, \ldots, m-1)$ is defined as the set of words $\sigma_1 \ldots \sigma_n$ from B_n such that

$$\sum_{i=1}^{n} \sigma_i i \equiv a \pmod{m}.$$

We shall show that each code $K_{n,m}^a$, for $m \geqslant n+1$, is a code correcting one deletion. Suppose that, as a result of one deletion, a word $x = \sigma_1 \ldots \sigma_n$ from $K_{n,m}^a$ has been transformed into the word $x' = \sigma'_1 \ldots \sigma'_{n-1}$. Then the number $|x'|$ and the least nonnegative residue of the number

$$a - \sum_{i=1}^{n-1} \sigma'_i i$$

modulo m, which we shall denote by a', may be regarded as known. In order to reconstruct the word x from the word x', it is evidently sufficient to know: 1) which of the binary symbols, 0 or 1, was deleted, and 2) either the number (which we shall denote by n_0) of zeros to the left of the deleted symbol, if this symbol is 1, or the number (which we shall denote by n_1) of ones to the right of

the deleted symbol, if this symbol is 0. But from the definition of the codes $K^a_{n,m}$ and the numbers n_0, n_1 it follows that, for $m \geqslant n+1$, either $a' = |x'| + 1 + n_0$ (if the symbol 1 was deleted), or $a' = n_1$ (if the symbol 0 was deleted), with $n_1 \leqslant |x'|$. Therefore, depending on whether a' is greater than the number $|x'|$ or not, one can determine which of the binary symbols was deleted, and then find the number n_0 or n_1, respectively. Consequently, by Lemma 1, each code $K^a_{n,m}$, for $m \geqslant n+1$, is a code correcting one deletion or insertion. Since each word from B_n belongs to one and only one of the m codes $K^a_{n,m}$ $(a = 0, 1, \ldots, m-1)$, at least one of these codes contains no fewer than $2^n/m$ words, which for $m = n+1$ gives the estimate (5).

2. Codes correcting deletions, insertions, and substitutions. We shall call a code K a **code correcting s deletions, insertions, and substitutions** if any binary word can be obtained from no more than one codeword of the code K by means of s or fewer deletions, insertions, and substitutions. It can be shown that the function $r(x, y)$, defined on pairs of binary words and equal to the least number of deletions, insertions, and substitutions transforming the word x into y, is a metric, and that a code K is a code correcting s deletions, insertions, and substitutions if and only if, for any distinct words x and y from K, one has $r(x, y) > 2s$. Denote by $M_s(n)$ the cardinality of the largest code in B_n correcting s deletions, insertions, and substitutions.

Theorem 2.

$$2^{n-1}/n \leqslant M_1(n) \leqslant 2^n/(n+1). \tag{6}$$

Proof. The upper bound is the Hamming bound ([4]) for codes correcting one substitution. To prove the lower bound, it is enough for us to show that all codes $K^a_{n,m}$ defined in the proof of Theorem 1, for $m \geqslant 2n$, are codes correcting one deletion, insertion, or substitution. The fact that these codes allow one to correct a deletion or insertion has already been proved. Further note that if, as a result of no more than one substitution, a word $\sigma_1 \ldots \sigma_n$ from $K^a_{n,m}$ has been transformed into the word $\sigma'_1 \ldots \sigma'_n$, then the minimum of the least nonnegative residues of the numbers

$$a - \sum_{i=1}^{n} \sigma'_i i \quad \text{and} \quad \sum_{i=1}^{n} \sigma'_i i - a$$

modulo $2n$ or more—

is equal to j, where j is the number of the substituted symbol (or $j = 0$, if no substitution occurred).

Using the method of proof of Lemma 2, one can establish that, for fixed s and $n \to \infty$,

$$\left((2s)!\Big/\sum_{i=0}^{s}2^{-i}C_{2s}^{2i}C_{2i}^{i}\right)\frac{2^n}{n^{2s}}\lesssim M_s(n)\lesssim s!\,\frac{2^n}{n^s}. \tag{7}$$

3. Use of codes in transmission (without synchronizing symbols) over channels with deletions, insertions, and substitutions. Denote by $l'_{s,n}$ ($l''_{s,n}$; $l_{s,n}$; $m_{s,n}$) a channel in which in each segment of length n there occur no more than s deletions (respectively insertions; deletions and insertions; deletions, insertions, and substitutions). We shall agree to write the sequence obtained at the output of the channel from an arbitrary infinite product $z_1z_2\ldots$ of words of the code J in the form $z'_1z'_2\ldots$, where z'_i denotes the word obtained from the codeword z_i as a result of errors in the channel. We shall call a code J **admissible** for the given channel if there exists a finite automaton* mapping any sequence $z'_1z'_2\ldots$ into the sequence $z_1z_2\ldots$. In order that a code J be admissible for the channels defined above, it is necessary (but, generally speaking, not sufficient) that it be a code correcting s errors of the corresponding kinds. In constructing admissible codes the following assertion is useful: for any binary words α and β, the codes K and $K_{\alpha,\beta}=\{\alpha x\beta,\ x\in K\}$ are codes correcting one and the same number of errors of the kinds under consideration. This assertion follows from the evident equalities $\rho(\alpha x\beta,\alpha y\beta)=\rho(x,y)$, $r(\alpha x\beta,\alpha y\beta)=r(x,y)$. In what follows the word $\beta\alpha$ plays the role of a comma between codewords, although it will, generally speaking, be distorted by errors in the channel.

Let us next note the important circumstance that, in contrast to the channel l_{sn}, in the case of the channels $l'_{s,n}$, $l_{s,n}$, $m_{s,n}$ for $s\geq 2$ (i.e., channels with two or more insertions) no code J makes it possible, from an arbitrary sequence $z'_1z'_2\ldots$, to determine where the word z'_1 ends. This leads to the fact that, in the indicated cases, in decoding it is necessary to proceed from the assumption that not only channel errors are possible, but also errors caused by an incorrect determination of the beginning of the next word z'_i (decoding errors). The idea of the constructions proposed below for the indicated channels is that, as a result of considering decoding errors as channel errors, each codeword should have to account for no more than s errors. This is achieved by a certain reduction in the length of the code and an appropriate choice of the comma $\beta\alpha$. The following assertions hold: 1) if a code K from B_{n-2s-1} is a code correcting s deletions, then the code $J=K_{1^s,0^s}$ is admissible for the channel $l'_{s,n}$; 2) if a code K from B_{n-4s} is a code correcting s insertions, then the code $J=K_{\Lambda,\,1^s0^s}$ is admissible for the channel $l''_{s,n}$; 3) if a code K from $B_{n-4(s+1)^2-2s}$ is a code correcting s deletions, insertions, and substitutions (deletions and insertions), then the code $J=K_{\Lambda,\,(1^{s+1}0^{s+1})^{s+1}1^s}$ is admissible** for the channel $m_{s,n}$ (respectively $l_{s,n}$).

Received
2 I 1965

CITED LITERATURE

[1] F. F. Sellers, Jr., IRE Trans., IT-8, No. 1 (1962).
[2] W. Feller, *An Introduction to Probability Theory and Its Applications*, 1964.
[3] R. R. Varshamov, G. M. Tenengolts, *Avtomatika i telemekhanika*, 26, No. 2 (1965).
[4] R. W. Hamming, Bell Syst. Techn. J., 29, No. 2 (1950).
[5] V. I. Levenshtein, *Problems of Cybernetics*, vol. 11, 1964.

* In a certain generalized sense (see, for example, (5)).

** In the case of the channel $m_{1,n}$ one can show that if a code K from B_{n-7} corrects one deletion, insertion, or substitution (for example, $K = K_{n-7,\,2(n-7)}$), then the code $J = K_{11,\,01}$ is admissible.

Note: Figure translations are in progress. See original paper for figures.