ON THE COMPLEXITY OF CODING

S. I. GELFAND, R. L. DOBRUSHIN and M. S. PINSKER
USSR

ABSTRACT

In the present work the task of construction of a scheme with a minimal number of binary summators is considered allowing to realize a linear binary block code, with optimal code distance. It is shown that a coding may be constructed with code distance of the order of dn, where $d < d_{V.G}$, n is the length of the code and $d_{V.G}n$, the asymptotic Varshamov–Hilbert bound, may be realized by schemes, containing approximately $c(d)n$ summators.

1. INTRODUCTION

By a scheme on summators with m inputs and n outputs we mean a directed graph G without cycles of the following form. In G we choose m nodes $a_1, \ldots, a_m$ called inputs, and n nodes $b_1, \ldots, b_n$, called outputs. The nodes $b_1, \ldots, b_n$ have no outgoing edges. Each node, except the nodes $a_1, \ldots, a_m$ have precisely two incoming edges, the nodes $a_1, \ldots, a_m$ have no incoming edges.

Definition 1. The complexity $h(G)$ of the scheme G is the number of nodes in G.

Let now be given a binary vector $x = (x_1, \ldots, x_m)$; $x_i = 0$ or 1. A state f of the scheme G, corresponding to the incoming vector x, is the assignment of a number $f(a)$, $f(a) = 0$ or 1, to every node a of the scheme G in such a way that the following conditions are fulfilled.

1) $f(a_i) = x_i$

2) If two edges r' and r'', come into the node a, r' coming out of a' and r'' out of a'', then* $f(a) = f(a') \oplus f(a'')$.

The following lemma is easily proved.

Lemma 1. For any G and x the state f exists and is unique.

Definition 2. Let G be a scheme with m inputs and n outputs, and $x = (x_1, \ldots, x_m)$ the incoming vector. Let f be the state of G, corresponding to x. We put $y_i = f(b_i)$, $i = \overline{1, n}$. The vector $y = (y_1, \ldots, y_n)$ is called the image of the vector x under the action of G, and is denoted by $G(x)$.

* Here and further, $\oplus$ designates the addition of numbers or binary vectors modulo 2.

It is easy to verify that $G(x' \oplus x'') = G(x') \oplus G(x'')$. Therefore the mapping $x \to y = G(x)$, where $x = (x_1, \ldots, x_m)$ gives some binary linear (n, m) code $\mathfrak{A}$. In such a situation we say that the scheme G realizes the code $\mathfrak{A}$. The rate of transmission R of such a code clearly does not exceed m/n, and $R < m/n$ if and only if* $G(x) = O$ for some $x \neq O$.

Definition 3. Let $\mathfrak{A}$ be a linear binary code; $G(\mathfrak{A})$ denotes the set of all schemes, realizing the code $\mathfrak{A}$.

Definition 4. $h(\mathfrak{A}) = \min_{G \in G(\mathfrak{A})} h(G)$.

2. THE MAIN THEOREM

We introduce the function** $H(x) = -x \log x - (1 - x) \log (1 - x)$. For every R, $0 < R < 1$, $d = d_{V.G}(R)$ denotes the smallest root of the equation

$$H(d) = 1 - R .$$

Then it is well known [2], that $d_{V.G}(R)n$ is the asymptotic lower Varshamov–Hilbert bound for the code distance of block codes of length n with rate R. More precisely, for every $d < d_{V.G}$ there exists for sufficiently large n a linear $(n, [Rn])$-code $\mathfrak{A}$ with code distance larger than dn.

Definition 5. Let R and d be given, $0 < d < d_{V.G}(R)$. We put $h(R, dn) = \min h(\mathfrak{A})$. Here the minimum is taken over the set of all linear binary $(n, [Rn])$ codes $\mathfrak{A}$ with code distance not smaller than dn.

The main theorem of our article establishes the asymptotic behaviour of $h(R, dn)$ for fixed R and d and growing n.

Theorem 1. There exist such c_1 and c_2, not depending on n, that

$$c_1 n < h(R, dn) < c_2 n .$$

Here, $c_1 \geq 1 + R + dR$. The dependence of c_2 on R and d is more complicated. (Cf. the remark at the end of section 4).

The proof of the upper bound in Theorem 1 will be given in the next section. Here we give the proof for the lower bound. For that purpose we show that, if scheme realizes an $(n, [Rn])$-code $\mathfrak{A}$ with $d(\mathfrak{A}) \geq dn$, then $h(G) \geq n + Rn + dRn$. Firstly it is clear that any scheme realizing an $(n, [Rn])$-code must have $[Rn]$ inputs and n outputs. Further, it is not difficult to show (cf. [1]) that, if the code distance of a code $\mathfrak{A}$, realized by the scheme G, is not smaller than dn, then at least one output is connected with at leats dRn inputs, which means that the scheme G has at least dRn nodes different from inputs and outputs. This concludes the proof of the lower bound in Theorem 1.

* O designates the vector $(0, \ldots, 0)$.
** Log designated the logarithm with basis 2.

3. THE UPPER BOUND

Everywhere in the following it will be handy for us to consider the case, when $Rn = m$ is integer. The passage to the general case does not lead to additional difficulties.

For the proof of the upper bound in Theorem 1 we need some auxiliary statements.

Proposition 1. There exist such constants α, $0 < \alpha < 1$ and c', that for any sufficiently large n a scheme G' may be constructed with $2m$ inputs and $m' \leq m$ outputs, having the following properties.

1) $G'(x) \neq O$ for all such vectors x that* $0 < w(x) < \alpha n$
2) $h(G') < c'n$.

Proposition 2. Let $\beta > 0$ be given. Then there exists a constant c'' with the following property. For any sufficiently large n, any m_1, $2Rn < m_1 < n + 2Rn$ and any set** Z, $|Z| \leq 2^{2m}$, of vectors with length m_1 such that $w(z) \geq \beta n$ for all $z \in Z$, there exists a scheme G'' with m_1 inputs and $2n$ outputs, for which $w(G''(z)) \geq 2dn$ and $h(G'') \leq c''n$. Here c'' depends on β, d, R, but not on n, m_1 and Z.

The proof of Propositions 1 and 2 will be given in Section 4.

Lemma 2. Let G_0 be an arbitrary scheme with m inputs and n outputs, and let $m' \leq m$. Then there exists a scheme G_1 with m' inputs and $n' \leq n$ outputs, having the following two properties.

1) $h(G_1) \leq h(G_0)$
2) Let $x = (x_1, \ldots x_m)$ be such a vector that $x_{m+1} = \ldots = x_m = 0$; $x' = (x_1, \ldots, x_m)$ is the "truncation" of the vector x. Let further $y = G_0(x)$, $y' = G_1(x'')$. Then $w(G_0(x)) = w(G_1(x'))$.

Proof. Let $a_1^0, \ldots, a_m^0$ be the inputs of G_0, and $b_1^0, \ldots, b_n^0$ the outputs of G_0. We shall construct the scheme G_1 in the following way. We remove from G_0 all its input nodes $a_{m'+1}^0, \ldots, a_m^0$ together with all of their outgoing edges. After that we may have nodes a with no incoming edges or nodes a' with only one incoming edge. In the first case we remove this node a from the scheme together with all its outgoing edges. In the second case we identify a' with the node of the scheme from which starts the only edge of a^0, which ends in a'. Repeating this process several times, we construct a new scheme G_1, considering those outputs $b_{i_1}^0, \ldots, b_{i_t}^0$ of the scheme G_0 as outputs of G_1 which have not been removed.

The proof of the fact that the scheme constructed in such a way has Properties 1 and 2, is rather easy and is left to the reader.

We shall give the proof of the upper bound in Theorem 1 by induction. Namely we assume that we have a scheme G_0 of complexity $h(G_0)$ with $m = Rn$ inputs $a_1^0, \ldots, a_m^0$ and n outputs $b_1^0, \ldots, b_n^0$ realizing the (n, m)-

* $w(x)$ is the Hamming weight of the vector x, i.e. the number of components of x, different from 0.
** $|Z|$ is the number of elements in the finite set Z.

code $\mathfrak{A}_0$ with $d(\mathfrak{A}_0) \geq dn$. Using this scheme we construct a scheme G with $2m$ inputs and $2n$ outputs, realizing the $(2n, 2m)$-code $\mathfrak{A}$ with $d(\mathfrak{A}) \geqslant 2dn$, and estimate the complexity of G. For the construction of G we need the auxiliary schemes G' and G'', which are mentioned in Propositions 1 and 2.

We will consider the inputs $a_1, \ldots, a_{2m}$ of the scheme G' as inputs of G. Further, using the scheme G_0, we construct a scheme G_1, satisfying the condition of Lemma 2, and identify the m' output nodes of G' with m' inputs of the scheme G_1. We construct further a scheme, satisfying the conditions of Proposition 2 with $m_1 = 2m + n'$, $\beta = \min(\alpha, d)$ and the set Z, which we define in the following way.

Let $x = (x_1, \ldots, x_{2m}) \neq 0$. We put $y = G_1(G'(x))$ and $z = (x, y)$, where z is a binary vector of length $n' + 2m$. We denote by Z the set of all vectors z obtained in such a way. It is clear that $|Z| = 2^{2m} - 1$. We remark that if $w(x) < \alpha n$ then by Proposition 1 and Lemma 2, $w(y) \geq dn$. Therefore $w(z) \geq n.\min(\alpha, d)$ for all $z \in Z$. Let $a''_1, \ldots, a''_{m_1}$ be the inputs of G''. We identify $a''_1, \ldots, a''_{2m}$ with $a'_1, \ldots, a'_{2m}$ and $a''_{2m+1}, \ldots, a''_{m_1}$ with the outputs $b_1, \ldots, b_{n'}$ of the scheme G_1. The outputs $b''_1, \ldots, b''_{2n}$ of the scheme G'' will be considered as the outputs of the scheme G.

We show that the scheme G has the required properties. First of all the fact that $w(z) \geq \beta n$ for all $z \in Z$ and Proposition 2 imply that $w(G(x)) \geq 2dn$ for all $x = (x_1, \ldots, x_{2m}) \neq 0$. Further it is clear that $h(G) = h(G') + h(G'') + h(G_1) - Rn - n \leq (c' + c'' - R - 1)n + h(G_0) = c_3 n + h(G_0)$ where c_3 is a constant, not depending on n.

Application of the method of mathematical induction now allows to conclude the proof of the upper bound in Theorem 1, with $c_2 = 2c_3 + \varepsilon$ for any $\varepsilon > 0$.

4. THE PROOF OF THE PROPOSITIONS

In this section we will give the proof of Propositions 1 and 2. The methods of proof of both propositions coincide. Namely, we construct a (finite) set of schemes (one for each case), define on this set a probability distribution (a finite set of schemes with a probability distribution on it will further called an ensemble of schemes), and show that the probability of finding among the schemes of our set a scheme, satisfying the required conditions, is different from 0. It implies that there exists at least one scheme with the required properties. We remark that our proof is a pure existence proof and does not allow to construct explicitly the required scheme. The problem of the explicit construction of such schemes remains open and forms a very interesting and important task.

Before coming to the proof of Propositions 1 and 2, we introduce a useful auxiliary notion.

In the above notion of a scheme on summators in every intermediate node of such a scheme, there takes place an addition modulo 2 of quantities coming into this node along precisely two edges. Sometimes we shall have to add more than two such quantities modulo 2. For that purpose the notion of t-summator is introduced.

Definition 6. Let t be an integer. t-summator is a scheme Σ^t with t inputs and one output such that

1) $h(\Sigma^t) = 2t - 1$
2) $\Sigma^t(x) = x_1 \oplus \ldots \oplus x_t$ if $x = (x_1, \ldots, x_t)$.

Lemma 3. Σ^t exists for any integer t.

The construction of Σ^t is easily carried out.

Proof of Proposition 1. We fix an integer l, the precise value of which will be chosen later. We construct the ensemble $\mathcal{E} = \mathcal{E}_l$ as follows. We consider $2m = 2Rn$ nodes $a'_1, \ldots a_{2m}$. l edges $r_i^{(1)}, \ldots, r_i^{(l)}$ come from each node a'_i, $i = \overline{1, 2m}$. Let us further consider m nodes $b'_1, \ldots, b'_m$ and let each of the edges $r_i^{(j)}$, $i = \overline{1, 2m}, j = \overline{1, l}$ be associated with one of the nodes $b'_k, k = \overline{1, m}$. For every such association of edges $r_i^{(j)}$ and nodes b'_k we construct a scheme on summators G' as follows. The inputs of G' are the nodes $a'_1, \ldots, a'_{2m}$. If some node b'_k is not associated with any edge $r_i^{(j)}$, then we remove this node. If the node b'_k is associated with precisely one edge $r_i^{(j)}$ (it clearly comes from a'_i), then we identify b'_k with a'_i. Let us remark that thus different nodes b'_k may be identified with one and the same node a'_i, and then these nodes b'_k are identified with each other. If, further, the node b'_k is associated with t edges $r_{i_1}^{(j_1)}, \ldots, r_{i_t}^{(j_t)}$, then we identify b'_k with the output of a t-summator Σ^t_k, and the nodes $a'_{i_1}, \ldots, a'_{i_t}$ with the inputs of Σ^t_k. The ensemble $\mathcal{E}_l$ consists of all schemes G, obtained by such a construction for different ways of associating the edges $r_i^{(j)}$ with the nodes b'_k. The probability distribution on $\mathcal{E}_l$ is defined by the property that every edge $r_i^{(j)}$ is associated with each node b'_k with equal probability and independently of other edges.

From the construction it is clear that $h(G) \leq 2lm + 2m \leq 2Rn(l + 1)$ for all $G \in \mathcal{E}_l$. Let us now fix the incoming vector $x = (x_1, \ldots, x_{2m})$. Let $w = w(x)$. We consider the event

$$A_x = \{G(x) = 0\}$$

and give an upper bound for $Pr\{A_x\}$. The construction of the ensemble $\mathcal{E}_l$ implies at once that $Pr\{A_x\} = Pr\{A_{x'}\}$, if $w(x) = w(x')$. Therefore we may consider vector x_0 of the form $x_0 = (1, \ldots, \underline{1, 0,} \ldots, 0)$, $w(x_0) = w$. We call those edges $r_i^{(j)}$ distinguished, for which $i = \overline{1, w}$. It is clear that the number of distinguished edges is equal to $N = wl$.

Let us consider a scheme $G \in \mathcal{E}_l$. It is obtained, as described above, from some distributions of all edges $r_i^{(j)}$, $i = \overline{1, 2_m}$, $j = \overline{1, l}$, on the nodes b'_k. Let ξ_k be the number of distinguished edges, corresponding to the node b'_k. Then ξ_k, $k = \overline{1, m}$ is a random variable on $\mathcal{E}_l$, and $\xi_1 + \ldots + \xi_m = N$. It is clear then that

$$A_{x_0} = \bigcup_{\substack{n_1, \ldots, n_m - \text{even} \\ \Sigma n_i = N}} \{\xi_1 = n_1; \ldots; \xi_m = n_m\}.$$

Let us now introduce the event A'_{x_0} by

$$A'_{x_0} = \bigcup_{\substack{n_1,\ldots,n_m \\ \Sigma n_i = N \\ n_i \neq 1 \text{ for all } i=\overline{1,m}}} \{\xi_1 = n_1; \ldots ; \xi_m = n_m\}.$$

Clearly $A'_{x_0} \supset A_{x_0}$, so $Pr\{A_{x_0}\} \leq Pr\{A'_{x_0}\}$. Let us now find $Pr\{A'_{x_0}\}$.

Lemma 4. There exists such $\alpha_1 > 0$ that for all sufficiently large n

$$Pr\{A'_{x_0}\} < (C^w_{2m})^{-1} m^{-1}$$

for all w such that $0 \leq w < \alpha_1 m$.

Proof. The event A'_{x_0} signifies that all ξ_k are either equal to 0, or larger than 1. Since $\Sigma\, \xi_k = N = wl$, the number of ξ_k is different from 0, it does not exceed $wl/2$. Since all ξ_k are equally distributed, it follows that

$$Pr\{A'_{x_0}\} \leq C^{wl/2}_m Pr\{\xi_{\frac{wl}{2}+1} = \xi_{\frac{wl}{2}+2} = \ldots = \xi_m = 0\}.$$

From the definition of the probability distribution on the ensemble $\mathcal{E}_l$ follows that the last probability equals

$$Pr\{\xi_{\frac{wl}{2}+1} = \ldots = \xi_m = 0\} = \left(\frac{wl}{2m}\right)^{wl}.$$

Therefore, for the proof of the lemma we have to estimate from above the quantity

$$T = m C^w_{2m} C^{\frac{wl}{2}}_m \left(\frac{wl}{2m}\right)^{wl}$$

We may assume that $w < m$ and $\frac{wl}{2} < \frac{m}{2}$. Under these conditions we have the inequalities

$$C^w_{2m} < c_4\, 2^{2mH(w/2m)}; \quad C^{\frac{wl}{2}}_m \leq c_5\, 2^{mH\left(\frac{wl}{2m}\right)}$$

where c_4, c_5 do not depend on w and m. Therefore, putting $\omega = w/m$, we have

$$T \leq c_6\, 2^{\log m + m\{2H(w/2) + H(wl/2) + wl \log wl/2\}}.$$

Further, since $w \geq 1$, $\log m \leq w \log m = w \log \omega - w \log w \leq - w \log \omega$. Moreover, there exists such ω_0 that for $\omega < \omega_0$ we have $H(\omega) \leq -\frac{11}{10}\omega \log \omega$. Therefore, if $\omega < \omega_1$ for some ω_1 we have

$$T \leq c_6\, 2^{m\left(-\frac{11}{10}\omega \log \frac{\omega}{2} - \frac{11\omega l}{20} \log \frac{\omega l}{2} + \omega l \log \frac{\omega l}{2}\right) - w \log w} = c_6\, 2^{\left(l - \frac{11 l}{20} - \frac{21}{10}\right) w \log w + Bw}$$

where B is some constant. Let us now put $l = 5$. Then $l - \frac{11}{20} l - \frac{21}{10} > \frac{1}{10}$ and $T \leq 2^{w\left(\frac{\log \omega}{10} + B\right) + \log c_6}$.

This formula implies that there exists such an $\alpha_1 < 1$ that for $1 < w = \omega n \leq \alpha_1 m$ the expression in the exponent is negative. For such w we have $T < 1$, and Lemma 4 is proved.

Now we can easily conclude the proof of Proposition 1. Indeed, let us put $\alpha_1 = \alpha_1 R$ (α_1 as in Lemma 4) and let A designate the event

$$A = \{G(x) = O \text{ for some } x,\ 1 \leq w(x) \leq \alpha n\}.$$

Then $A = \bigcup_{x, 1 \leq w(x) \leq \alpha n} A_x$, so

$$Pr\{A\} \leq \sum_{1 \leq w(x) \leq \alpha n} Pr\{A_x\} \leq \sum_{1 \leq w \leq \alpha n} Pr\{A'_{x_0}\} \leq m^{-1} \sum_{w=1}^{\alpha n} 1 < 1$$

since $\alpha n < m$. This means that in the ensemble $\mathcal{E}_l$ for $l = 5$ there exists at least one scheme, satisfying the conditions of Proposition 1.

Proof of Proposition 2. Let us remind that with a given set Z, $|Z| \leq 2^{2m}$ of vectors of length $m_1 \leq n(2R + 1)$ such that $w(z) \geq \beta n$ for $z \in Z$, we have to construct a scheme G'', with m_1 inputs and $2n$ outputs, such that $w(G(z)) \geq 2dn$ for all $z \in Z$ and $h(G'') \leq c'' n$. For this purpose it is sufficient to construct an ensemble $\mathcal{E}$ of schemes with m_1 inputs and $2n$ outputs, satisfying the following conditions.

a) $h(G) \leq c' n$ for all $G \in \mathcal{E}$

b) $Pr\{w(G(x)) < 2dn\} < 2^{-2m}$

for any fixed $z = (z_1, \ldots, z_{m_1})$ with $w(z) > \beta n$.

We will construct this ensemble in the following way. Let us fix an odd $t > 0$. (The precise value of t will depend on β and d). Every scheme G of the ensemble $\mathcal{E}$ has $2n$ outputs, $b_1, \ldots, b_{2n}$, and every output b_k is the output of a t-summator Σ^t_k. The inputs $a^{(j)}_k$, $k = \overline{1, 2n}$, $j = \overline{1, t}$ of all summators Σ^t_k (there are totally $2nt$ of these) are identified with the inputs $a_1, \ldots, a_{m_1}$ of the scheme G. The probability assignment on $\mathcal{E}$ is defined by the condition that each of $2nt$ nodes $a^{(j)}_k$ is identified with one of the nodes $a_1, \ldots, a_{m_1}$ with equal probability and independently from the other nodes.

Clearly, $h(G) = m_1 + 2n(t - 1) \leq n(2R - 2t - 1)$ for all $G \in \mathcal{E}$. Therefore condition a) is fulfilled for the ensemble $\mathcal{E}$.

Let us now find $Pr\{w(G(z)) < 2dn\}$. Let an incoming word z, $w(z) = w$ be given. We call those nodes $a^{(j)}_k$ distinguished, which are identified with a node a_i with $z_i = 1$. Then the number of distinguished nodes between $a^{(1)}_k, \ldots, a^{(t)}_k$ is a random variable η_k on $\mathcal{E}$, and from the construction of $\mathcal{E}$ at once follows that the variables η_k are equally distributed, mutually independent and

$$Pr\{\eta_k = q\} = C^t_q \left(\frac{w}{m_1}\right)^q \left(1 - \frac{w}{m_1}\right)^{t-q}.$$

Let further $G(z) = y = (y_1, \ldots, y_{2n})$. We define the random variable ν_k, $k = \overline{1, 2n}$ on $\mathcal{E}$ by $\nu_k = y_k$. Then $\nu_k = 0$ if η_k is even and $\nu_k = 1$, if η_k is odd.

Therefore the variables ν_k are also equally distributed, mutually independent and

$$Pr\{\nu_k = 1\} = \sum_{\substack{q=1 \\ q=\text{odd}}}^{t} C_t^q \left(\frac{w}{m_1}\right)^q \left(1 - \frac{w}{m_1}\right)^{t-q} = \frac{1}{2} - \frac{1}{2}\left(1 - \frac{2w}{m_1}\right)^t = p_t\left(\frac{w}{m_1}\right).$$

Therefore,

$$Pr\{w(G(z)) \leq 2dn\} = Pr\left\{\sum_{k=1}^{2n} \nu_k \leq 2dn\right\} =$$

$$\sum_{i=0}^{2dn} C_{2n}^i \left[p_t\left(\frac{w}{m_1}\right)\right]^i \left[1 - p_t\left(\frac{w}{m_1}\right)\right]^{2n-i}.$$

For the last sum we have the following estimation (Chernov bound, [2])

$$\sum_{i=0}^{2dn} c_{2n}^i p^i (1-p)^{2n-1} \leq 2^{2n\{H(d)+d \log p + (1-d) \log (1-p)\}}$$

when $d < p$. Since $d < d_{V \cdot G} < 1/2$, we have $\theta = H(d) - 1 + R < 0$. Let us put $\beta_1 = \beta/(1 + 2R)$ and choose the odd t which was arbitrary until now so large that $d \log p_t(\beta_1) + (1-d) \log [1 - p_t(\beta_1)] < -1 - \varepsilon$ (this is possible, since $p_t(\beta_1) \to 1/2$ for $t \to \infty$). Since t is odd, $p_t(w/m_1) > p_t(\beta_1)$ for all $w > \beta n$ and $m_1 < n(1 + 2R)$. So

$$2^{H(d)+d \log p + (1-d) \log (1-p)} < 2^{-2nR}$$

and

$$Pr\{w(G(z)) < 2dn\} < 2^{-2nR}.$$

Thus, for the chosen t the ensemble $\mathcal{E}_t$ satisfies the conditions a) and b), which means that it contains a scheme G'', satisfying the conditions of Proposition 2.

Remark. The complexity of the scheme G'' in Proposition 2 depends on the number t, which is chosen as the smallest number, for which the inequality

$$d \log p_t(\beta_1) + (1-d) \log [1 - p_t(\beta_1)] < -H(d) - R$$

is fulfilled. In particular, for $d \to d_{V,G}$ the number t grows as $|\log (d_{V,G} - d)|$. Therefore the constant c_2 in Theorem 1 grows in the same way when $d \to d_{V,G}$. The question, if c_2 may be found, not depending on d, remains open.

REFERENCES

1. Gelfand, S. I. and Dobrushin, R. L., Construction of asymptotically optimal codes by scheme of constant depth. *Problems of Control and Information Theory* **1** (1972) *3–4*.
2. Peterson, W. W., *Error-correcting codes*. MIT Press, Cambridge, Mass.