В. И. Левенштейн, Двоичные коды с исправлением выпадений, вставок и замещений символов, *Докл. АН СССР*, 1965, том 163, номер 4, 845–848

Доклады Академии наук СССР
1965. Том 163, № 4

КИБЕРНЕТИКА И ТЕОРИЯ РЕГУЛИРОВАНИЯ

В. И. ЛЕВЕНШТЕЙН

# ДВОИЧНЫЕ КОДЫ С ИСПРАВЛЕНИЕМ ВЫПАДЕНИЙ, ВСТАВОК И ЗАМЕЩЕНИЙ СИМВОЛОВ

*(Представлено академиком П. С. Новиковым 4 I 1965)*

При исследовании вопросов передачи двоичной информации по каналам обычно рассматривается модель канала, в которой допускаются сбои вида $0 \to 1, 1 \to 0$, называемые в дальнейшем замещениями. В настоящей работе (как и в ([1])) исследуется модель канала, в которой допускаются также сбои вида $0 \to \Lambda$, $1 \to \Lambda$, называемые выпадениями, и сбои вида $\Lambda \to 0$, $\Lambda \to 1$, называемые вставками (здесь $\Lambda$ — пустое слово). Для таких каналов по аналогии с комбинаторной задачей построения оптимальных кодов с исправлением $s$ замещений рассматриваются задачи построения оптимальных кодов с исправлением выпадений, вставок и замещений.

1. Коды с исправлением выпадений и вставок. Слова в алфавите $\{0, 1\}$ будем называть двоичными словами. Произвольное множество двоичных слов фиксированной длины будем называть кодом*. Код $K$ назовем кодом с исправлением $s$ выпадений (кодом с исправлением $s$ вставок), если любое двоичное слово может быть получено не более чем из одного слова кода $K$ путем $s$ или менее выпадений (соответственно вставок). Код $K$ назовем кодом с исправлением $s$ выпадений и вставок, если любое двоичное слово может быть получено не более чем из одного слова кода $K$ путем $s$ или менее выпадений и вставок. Последнее свойство обеспечивает возможность однозначного определения исходного кодового слова по слову, полученному из него некоторым числом $i$ ($i \geqslant 0$) выпадений и некоторым числом $j$ ($j \geqslant 0$) вставок, если $i + j \leqslant s$. Следующее утверждение показывает, что все приведенные выше определения кодов равносильны.

Лемма 1. *Любой код с исправлением $s$ выпадений (равно как и любой код с исправлением $s$ вставок) является кодом с исправлением $s$ выпадений и вставок.*

Доказательство (от противного). Пусть из слова $x$ длины $n$ путем $i_1$ выпадений и $j_1$ вставок, где $i_1 + j_1 \leqslant s$, и из слова $y$ длины $n$ путем $i_2$ выпадений и $j_2$ вставок, где $i_2 + j_2 \leqslant s$, получается одно и то же слово $z$. Если в слове $z$ опустить (вставить) те символы, которые при получении $z$ вставляются (опускаются) хотя бы в одном из слов $x$ и $y$, то, как легко видеть, получится слово, которое можно получить и из $x$ и из $y$ не более чем $\max(i_2 + j_1, j_2 + i_1)$ выпадениями (соответственно вставками). Ввиду равенства длин слов $x$ и $y$ $j_1 - i_1 = j_2 - i_2$ и, следовательно, $i_2 + j_1 = j_2 + i_1 = \frac{1}{2}(i_1 + i_2 + j_1 + j_2) \leqslant s$, что доказывает лемму 1.

Коды с исправлением $s$ выпадений и вставок допускают другое, метрическое описание. Рассмотрим функцию $\rho(x, y)$, определенную на парах двоичных слов и равную наименьшему числу выпадений и вставок, преобразующих слово $x$ в $y$. Нетрудно показать, что функция $\rho(x, y)$ является метрикой, причем код $K$ является кодом с исправлением $s$ выпадений и вставок тогда и только тогда, когда для любых двух различных слов $x$ и $y$ из $K$ имеет место $\rho(x, y) > 2s$.

Пусть $B_n$ — множество всех двоичных слов длины $n$. Для произвольного слова $x$ из $B_n$ обозначим через $|x|$ число единиц в слове $x$, а через

* Дальнейшие определения имеют также смысл, если под кодом понимать произвольное множество слов (быть может, различной длины) в некотором алфавите из $r$ букв ($r \geqslant 2$). Отметим, однако, что в случае слов различной длины лемма 1, вообще говоря, уже не верна.

$\|x\|$ — число серий* слова $x$ и оценим число $P_s(x)$ (число $Q_s(x)$) различных слов, получаемых из $x$ путем $s$ выпадений (соответственно $s$ вставок). Имеют место оценки:

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

$$\sum_{i=0}^{s} C_n{}^i 2^{s-i} \leqslant Q_s(x) \leqslant \sum_{i=0}^{s} C_n{}^i C_s{}^i 2^{s-i}. \quad (2)$$

Для доказательства верхней оценки в (1) заметим, что каждое слово, получаемое из $x$ выпадениями, однозначно определяется указанием числа символов, выпавших из каждой серии, и, следовательно, $P_s(x)$ не превышает числа упорядоченных разбиений числа $s$ на $\|x\|$ неотрицательных слагаемых. С другой стороны, легко видеть, что если в любых $s$ попарно несмежных сериях слова $x$ выбросить по одному символу, то все полученные таким образом слова будут различны. Это дает нижнюю оценку в (1), если заметить, что количество таких слов равно числу упорядоченных разбиений числа $\|x\| - s$ на $s+1$ неотрицательных слагаемых, лишь два из которых, быть может, нули. Верхняя оценка в (2) следует из того, что каждое слово, получаемое из слова $x = \sigma_1 \ldots \sigma_n$ путем $s$ вставок, может быть получено следующим образом. При некотором $i$ ($i = 0, 1, \ldots, s$) выбирается $i$ номеров $n_1, \ldots, n_i$ $(1 \leqslant n_1 < \ldots < n_i \leqslant n)$ и $i+1$ слов $\beta_1, \ldots, \beta_i, \beta_{i+1}$, сумма длин которых равна $s$, причем каждое из первых $i$ слов $\beta_j$ непусто и не оканчивается символом $\sigma_{n_j}$; затем каждое слово $\beta_j$ $(j = 1, \ldots, i)$ вставляется в слово $x$ перед символом $\sigma_{n_j}$, а слово $\beta_{i+1}$ после символа $\sigma_n$. Нижняя оценка в (2) следует из того, что если каждое из слов $\beta_1, \ldots, \beta_i$ имеет длину один, то все слова, получаемые из $x$ указанным выше способом, различны.

Отметим, что из (1) и (2) следует, что $P_1(x) = \|x\|$ и $Q_1(x) = n + 2$.

Обозначим через $L_s(n)$ мощность (число слов) максимального в $B_n$ кода с исправлением $s$ выпадений и вставок.

Лемма 2**. *При фиксированном $s$ и $n \to \infty$*

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

Доказательство. Пусть $K$ — максимальный в $B_n$ код с исправлением $s$ выпадений и вставок и пусть при произвольном $k$ $(1 \leqslant k < n/2)$ $L_s(n) = L_k' + L_k''$, где $L_k'$ — число слов $x$ кода $K$ таких, что $k < \|x\| < n - k$. Из определения кода $K$ следует, что $\sum_{x \in K} P_s(x) \leqslant 2^{n-s}$, а из его максимальности $\sum_{x \in K} R_{2s}(x) \geqslant 2^n$, где $R_{2s}(x)$ — число слов, находящихся на расстоянии $2s$ или менее (в метрике $\rho(x, y)$) от слова $x$. Используя (1) и (2), получаем $2^{n-s} \geqslant L_k' C_{k-s}{}^s$ и

$$2^n \leqslant (L_k' C^s_{n-k+s} + L_k'' C^s_{n+s-1}) \sum_{i=0}^{s} C^i_{n-1} C_s{}^i 2^{s-i}.$$

Из последних неравенств следуют оценки (3), если заметить, что $L_k'' \leqslant 2\left(\sum_{i=1}^{k} C^{i-1}_{n-1} + \sum_{i=n-k}^{n} C^{i-1}_{n-1}\right) = 2 \sum_{i=0}^{k} C^i_n$ (так как число слов из $B_n$, имеющих $i$ серий, равно $2C^{i-1}_{n-1}$), и воспользоваться тем, что $\sum_{i=0}^{k} C_n{}^i = o\left(\frac{2^n}{n^{2s}}\right)$ при $k = [n/2 - \sqrt{sn \ln n}]$ и $n \to \infty$ (см., например, ([2])).

* Серией слова $x$ называется максимальное подслово слова $x$, состоящее из одинаковых символов. Например, слово $x = 01101$ имеет 4 серии.

** В дальнейшем запись $f(n) \lesssim g(n)$ означает, что $\overline{\lim_{n \to \infty}} f(n) / g(n) \leqslant 1$, а запись $f(n) \sim g(n)$ означает, что $\lim_{n \to \infty} f(n) / g(n) = 1$.

Теорема 1.

$$L_1(n) \sim 2^n / n. \qquad (4)$$

Доказательство. В силу леммы 2, нам достаточно показать, что

$$L_1(n) \geqslant 2^n / (n+1). \qquad (5)$$

Чтобы доказать это, воспользуемся одной конструкцией Варшамова — Тененгольца [3]. Рассмотрим класс кодов $K^a_{n,m}$, где каждый $K^a_{n,m}$ ($a = 0, 1, \ldots, m-1$) определяется как множество слов $\sigma_1 \ldots \sigma_n$ из $B_n$ таких, что $\sum\limits_{i=1}^{n} \sigma_i i \equiv a (\mathrm{mod}\, m)$. Покажем, что каждый код $K^a_{n,m}$ при $m \geqslant n+1$ является кодом с исправлением одного выпадения. Пусть в результате одного выпадения слово $x = \sigma_1 \ldots \sigma_n$ из $K^a_{n,m}$ преобразовалось в слово $x' = \sigma_1' \ldots \sigma'_{n-1}$. Тогда можно считать известными число $|x'|$ и наименьший неотрицательный вычет числа $a - \sum\limits_{i=1}^{n-1} \sigma'_i i$ по $\mathrm{mod}\, m$, который мы обозначим через $a'$. Для того чтобы восстановить слово $x$ по слову $x'$, очевидно, достаточно знать: 1) какой из двоичных символов 0 или 1 выпал и 2) либо число (которое мы обозначим через $n_0$) нулей левее выпавшего символа, если этот символ есть 1, либо число (которое мы обозначим через $n_1$) единиц правее выпавшего символа, если этот символ есть 0. Но из определения кодов $K^a_{n,m}$ и чисел $n_0$, $n_1$ следует, что при $m \geqslant n+1$ либо $a' = |x'| + 1 + n_0$ (если выпал символ 1), либо $a' = n_1$ (если выпал символ 0), причем $n_1 \leqslant |x'|$. Поэтому в зависимости от того, больше $a'$ числа $|x'|$ или нет, можно определить, какой из двоичных символов выпал, а затем найти число $n_0$ или $n_1$ соответственно. Следовательно, по лемме 1, каждый код $K^a_{n,m}$ при $m \geqslant n+1$ является кодом с исправлением одного выпадения или вставки. Поскольку каждое слово из $B_n$ принадлежит одному и только одному из $m$ кодов $K^a_{n,m}$ ($a = 0, 1, \ldots, m-1$), то по крайней мере один из этих кодов содержит не менее $2^n / m$ слов, что при $m = n+1$ дает оценку (5).

2. Коды с исправлением выпадений, вставок и замещений. Код $K$ назовем кодом с исправлением $s$ выпадений, вставок и замещений, если любое двоичное слово может быть получено не более чем из одного слова кода $K$ путем $s$ или менее выпадений, вставок и замещений. Можно показать, что функция $r(x,y)$, определенная на парах двоичных слов и равная наименьшему числу выпадений, вставок и замещений, преобразующих слово $x$ в $y$, является метрикой, причем код $K$ является кодом с исправлением $s$ выпадений, вставок и замещений тогда и только тогда, когда для любых различных слов $x$ и $y$ из $K$ имеет место $r(x, y) > 2s$. Обозначим через $M_s(n)$ мощность максимального в $B_n$ кода с исправлением $s$ выпадений, вставок и замещений.

Теорема 2.

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

Доказательство. Верхняя оценка — это оценка Хемминга [4] для кодов с исправлением одного замещения. Для доказательства нижней оценки нам достаточно показать, что все коды $K^a_{n,m}$, определенные при доказательстве теоремы 1, при $m \geqslant 2n$ являются кодами с исправлением одного выпадения, вставки или замещения. Тот факт, что эти коды позволяют исправить выпадение или вставку, уже доказан. Заметим далее, что если в результате не более одного замещения слово $\sigma_1 \ldots \sigma_n$ из $K^a_{n,m}$ преобразовалось в слово $\sigma_1' \ldots \sigma_n'$, то минимальный из наименьших неотрицательных вычетов чисел $a - \sum\limits_{i=1}^{n} \sigma_i' i$ и $\sum\limits_{i=1}^{n} \sigma_i' i - a$ по модулю $2n$ или более равен $j$, где $j$ — номер замещенного символа (или $j = 0$, если замещения не произошло).

Используя метод доказательства леммы 2, можно установить, что при фиксированном $s$ и $n \to \infty$

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

3. Использование кодов при передаче (без синхронизирующих символов) по каналам с выпадениями, вставками и замещениями. Обозначим через $l'_{s,n}$ ($l''_{s,n}$; $l_{s,n}$; $m_{s,n}$) канал, в котором в каждом отрезке длины $n$ происходит не более $s$ выпадений (соответственно вставок; выпадений и вставок; выпадений, вставок и замещений). Условимся последовательность, полученную на выходе канала из произвольного бесконечного произведения $z_1 z_2 \ldots$ слов кода $J$, записывать в виде $z_1' z_2' \ldots$, где через $z_i'$ обозначено слово, полученное из кодового слова $z_i$ в результате сбоев в канале. Код $J$ будем называть **допустимым** для данного канала, если существует конечный автомат*, отображающий любую последовательность $z_1' z_2' \ldots$ в последовательность $z_1 z_2 \ldots$ Для того чтобы код $J$ был допустимым для определенных выше каналов, необходимо (но, вообще говоря, не достаточно), чтобы он был кодом с исправлением $s$ сбоев соответствующих видов. При построении допустимых кодов полезно следующее утверждение: для любых двоичных слов $\alpha$ и $\beta$ коды $K$ и $K_{\alpha,\beta} = \{\alpha x \beta,\ x \in K\}$ являются кодами с исправлением одного и того же числа сбоев рассматриваемых видов. Это утверждение следует из очевидных равенств $\rho(\alpha x \beta, \alpha y \beta) = \rho(x, y)$, $r(\alpha x \beta, \alpha y \beta) = r(x, y)$. В дальнейшем слово $\beta\alpha$ играет роль запятой между кодовыми словами, хотя оно, вообще говоря, будет искажаться сбоями в канале.

Отметим далее то важное обстоятельство, что, в отличие от канала $l_{sn}'$, *в случае каналов* $l''_{s,n}$, $l_{s,n}$, $m_{s,n}$ *при* $s \geqslant 2$ *(т. е. каналов с двумя или более вставками) никакой код* $J$ *не позволяет по любой последовательности* $z_1' z_2' \ldots$ *определить, где оканчивается слово* $z_1'$. Это приводит к тому, что в указанных случаях при декодировании необходимо исходить из того, что возможны не только сбои канала, но и сбои, вызванные неправильным определением начала очередного слова $z_i'$ (сбои декодирования). Идея предложенных ниже конструкций для указанных каналов состоит в том, *чтобы в результате рассмотрения сбоев декодирования как сбоев канала на каждое кодовое слово приходилось не более* $s$ *сбоев*. Это достигается за счет некоторого уменьшения длины кода и подходящего выбора запятой $\beta\alpha$. Справедливы следующие утверждения: 1) *если код* $K$ *из* $B_{n-2s-1}$ *является кодом с исправлением* $s$ *выпадений, то код* $J = K_{1^s, 0^s}$ *допустим для канала* $l'_{s,n}$; 2) *если код* $K$ *из* $B_{n-4s}$ *является кодом с исправлением* $s$ *вставок, то код* $J = K_{\Lambda, 1^s 0^s}$ *допустим для канала* $l''_{s,n}$; 3) *если код* $K$ *из* $B_{n-4(s+1)^2-2s}$ *является кодом с исправлением* $s$ *выпадений, вставок и замещений (выпадений и вставок), то код* $J = K_{\Lambda, (1^{s+1}0^{s+1})^{s+1}1^s}$ *допустим*** *для канала* $m_{s,n}$ *(соответственно* $l_{s,n}$*)*.

Поступило
2 I 1965

## ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] F. F. Sellers Jr., IRE Trans., IT-8, № 1 (1962). [2] В. Феллер, Введение в теорию вероятностей и ее приложения, 1964. [3] Р. Р. Варшамов, Г. М. Тененгольц, Автоматика и телемеханика, 26, № 2 (1965). [4] R. W. Hamming, Bell Syst. Techn. J., 29, № 2 (1950). [5] В. И. Левенштейн, Проблемы кибернетики, в. 11, 1964.

* В некотором обобщенном смысле (см., например, ([5])).

** В случае канала $m_{1,n}$ можно показать, что если код $K$ из $B_{n-7}$ исправляет одно выпадение, вставку или замещение (например, $K = K_{n-7, 2(n-7)}$), то код $J = K_{11, 01}$ является допустимым.