COMMA-FREE CODES

S. W. GOLOMB, BASIL GORDON AND L. R. WELCH

1. A General Combinatorial Problem. Let n be a fixed positive integer, and consider an alphabet consisting of the numbers 1, 2, . . . , n. With this alphabet form all possible k-letter words $(a_1\ a_2\ \ldots\ a_k)$, where k is also fixed. There are evidently n^k such words in all.

Definition: A set D of k-letter words is called a *comma-free dictionary* if whenever $(a_1\ a_2\ \ldots\ a_k)$ and $(b_1\ b_2\ \ldots\ b_k)$ are in D, the "overlaps" $(a_2\ a_3\ \ldots\ a_k\ b_1)$, $(a_3\ \ldots\ a_k\ b_1\ b_2)$, . . . , $(a_k\ b_1\ \ldots\ b_{k-1})$ are not in D.

The problem to be investigated here is that of determining the greatest number of words that a comma-free dictionary can possess. We denote this number by $W_k(n)$.

THEOREM 1.

$$W_k(n) \leqslant \frac{1}{k} \sum \mu(d)\, n^{k/d},$$

where the summation is extended over all divisors d of k, and $\mu(d)$ is the Möbius function, defined by

$$\mu(d) = \begin{cases} 1 & \text{if } d = 1 \\ 0 & \text{if } d \text{ has any square factor} \\ (-1)^r & \text{if } d = p_1\, p_2 \ldots p_r, \text{ where } p_1, \ldots, p_r \text{ are distinct primes.} \end{cases}$$

Proof. Let d be a divisor of k. We say that a word $(a_1\ a_2\ \ldots\ a_k)$ has subperiod d if it is of the form $(a_1\ a_2\ \ldots\ a_d\ a_1\ a_2\ \ldots\ a_d\ \ldots\ a_1\ a_2\ \ldots\ a_d)$, and if d is the smallest number for which this is true. For example, if $k = 6$, then $(a\ a\ a\ a\ a\ a)$ has subperiod 1, $(a\ b\ a\ b\ a\ b)$ has subperiod 2 if $a \neq b$, $(a\ b\ c\ a\ b\ c)$ has subperiod 3 if $a \neq b$ or $b \neq c$, and all other words have subperiod 6. Any word w of a comma-free dictionary must have subperiod k because otherwise ww would contain an overlap of w. (Consider for example, $[a\ b\ c\ a\ b\ c]\ [a\ b\ c\ a\ b\ c]$.) We shall call words of subperiod k *primitive*.

For later purposes it is convenient to call two words *equivalent* if one is a cyclic permutation of the other, and to speak of $(a_1\ a_2\ \ldots\ a_k)$, $(a_2\ \ldots\ a_k\ a_1)$, . . . , $(a_k\ a_1\ \ldots\ a_{k-1})$ as forming an equivalence class. If $(a_1\ a_2\ \ldots\ a_k)$ is primitive, then its equivalence class is also called primitive, and consists of k distinct words. At most one of these can be a word in D, for otherwise a contradiction

Received July 11, 1957.

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

would again arise upon considering the overlaps of ww. Hence, if $P_k(n)$ is the total number of primitive words, then

$$W_k(n) \leqslant \frac{1}{k} P_k(n).$$

But since each of the n^k words has some subperiod, $P_k(n)$ satisfies the equation

$$\sum_{d/k} P_d(n) = n^k,$$

from which we obtain

$$P_k(n) = \sum_{d/k} \mu(d)\, n^{k/d}$$

by Möbius inversion. For the Möbius inversion formula, cf. **(3**, p. 28**)**.

2. Results for k odd. Theorem 1 gives a general upper bound for $W_k(n)$, of which a few examples are

$$W_1(n) \leqslant n, \quad W_2(n) \leqslant \tfrac{1}{2}(n^2 - n), \quad W_3(n) \leqslant \tfrac{1}{3}(n^3 - n), \quad W_4(n) \leqslant \tfrac{1}{4}(n^4 - n^2).$$

In many cases this upper bound is actually attained. We believe this to be true for all odd k, and have proved it for all odd $k \leqslant 15$. Note, from the proof of Theorem 1, that the upper bound will be attained if and only if a word can be chosen from each primitive equivalence class so as to form a comma-free dictionary.

THEOREM 2. *For arbitrary n,*

$$W_k(n) = \frac{1}{k} \sum_{d/k} \mu(d)\, n^{k/d}$$

if $k = 1, 3, 5, 7, 9, 11, 13, 15.$

Proof. For $k = 1$, the proof that $W_1(n) = n$ is immediate. For the other values of k we shall show how to select a word from each primitive equivalence class in such a way that a comma-free dictionary is obtained.

(i) In the case $k = 3$, let D be the set of all words $(a\ b\ c)$ satisfying the inequalities $a < b \geqslant c$. It is immediately seen that D is comma free. In order to show that the number of words in D is $\frac{1}{3}(n^3 - n)$, one could, of course, count the number of solutions of the inequalities $a < b \geqslant c$, where a, b, c are integers between 1 and n. But it is simpler to observe that if $(a_1\ a_2\ a_3)$ is any primitive word (that is, one for which $a_1 = a_2 = a_3$ does not hold), then some cyclic permutation of it clearly satisfies $a < b \geqslant c$. In particular $W_3(4) = 20$, a fact which will be useful in section 5.

(ii) For $k = 5$, the procedure is similar but more complex. Let D consist of all words $(a\ b\ c\ d\ e)$ satisfying $a < b \geqslant c$, $d \geqslant e$, and also of all words satisfying $a < b < c < d \geqslant e$. It can be readily verified that D is comma-free. In order to show that the number of elements in D is the upper bound (in this

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

case $\frac{1}{5}(n^5 - n)$), we must prove that every primitive equivalence class contains a word of D. For this purpose let $+$ denote any number which is > 0, and $-$ any number which is $\leqslant 0$. Using this notation the elements of D can be characterized as those words $(a\ b\ c\ d\ e)$ for which the sequence of differences $b - a$, $c - b$, $d - c$, $e - d$ is of one of the forms $+\ -\ -\ -$, $+\ -\ +\ -$, or $+\ +\ +\ -$. These patterns are precisely those which begin with an odd number of $+$'s and end with an odd number of $-$'s, a property which we shall call property P. (Incidentally, in the case $k = 3$ our dictionary consisted of words $[a\ b\ c]$ for which the differences $b - a$, $c - b$ were of the form $+\ -$, that is, possessed property P.) Given any primitive word $(p\ q\ r\ s\ t)$ we form the differences $q - p$, $r - q$, $s - r$, $t - s$, $p - t$ obtained by representing p, q, r, s, t as points on a circle. We call $p - t$ the improper difference. By performing a suitable cyclic permutation on $(p\ q\ r\ s\ t)$ we can arrange matters so that any one of the five differences becomes the improper one.

Now by primitivity, both $+$'s and $-$'s appear among the differences, and since the total number of signs is 5, there must occur someplace a run of $-$'s followed by a run of $+$'s, the lengths of these runs being of opposite parity (note that this result depends only on the fact that the total number of signs is *odd*). Permuting cyclically we can put the run of $+$'s at the beginning, the run of $-$'s at the end, and make the improper difference have the sign which occurred an even number of times. The proper differences will then satisfy property P, and hence, given any primitive word, some cyclic permutation of it is in D.

(iii) For $k = 7$ we use the same method. Every primitive word has some cyclic permutation with property P. Its proper differences will then have one of the following 8 patterns:

$$(+\ +\ +\ +\ +\ -)\ (+\ +\ +\ -\ +\ -)\ (+\ +\ +\ -\ -\ -)\ (+\ -\ +\ +\ +\ -)$$
$$(+\ -\ +\ -\ -\ -)\ (+\ -\ -\ +\ +\ -)\ (+\ -\ -\ -\ +\ -)\ (+\ -\ -\ -\ -\ -)$$

Letting D consist of all such words, we find that D is comma-free. (The verification begins to become tedious, but is straightforward. The first overlap of two words in D begins with an even number of $+$'s, hence is not in D, the second overlap ends with a $+$, etc.)

(iv) When $k = 9$, the difficulty arises that there may be more than one word in a primitive equivalence class with property P. This happens for words $(a_1\ a_2\ a_3\ a_4\ a_5\ a_6\ a_7\ a_8\ a_9)$ with $a_1 < a_2 \geqslant a_3$, $a_4 < a_5 \geqslant a_6$, $a_7 < a_8 \geqslant a_9$. Here the permutations $(a_4\ a_5\ a_6\ a_7\ a_8\ a_9\ a_1\ a_2\ a_3)$ and $(a_7\ a_8\ a_9\ a_1\ a_2\ a_3\ a_4\ a_5\ a_6)$ also have property P. But notice that these words consist of three blocks of three letters, each of the type used for $k = 3$. This suggests the idea of ordering the 3-letter words $(a\ b\ c)$ with $a < b \geqslant c$ in some fashion (say lexicographically), and choosing for the dictionary D that one of the three possibilities which is of the form $w_1 < w_2 \geqslant w_3$ in this ordering. For example, in the case of the word (1 3 1 1 2 2 2 3 1), the permutation (1 2 2 2 3 1 1 3 1) would be selected for D, because $(122) < (231) \geqslant (131)$ if lexicographic ordering is

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

employed. Adopting this convention, the dictionary which results is comma-free.

(v) For $k = 11$, 13, and 15 the same methods can be used, but the work becomes increasingly cumbersome. It is conceivable that all odd k can be treated in this manner, but we have stopped with the proof that $15W_{15}(n) = n^{15} - n^5 - n^3 + n$.

This case, the first where k has two distinct prime factors, is particularly powerful evidence for the validity of the general conjecture.

3. Results for Even k. When k is even, the results are much less complete, and we cannot even formulate a plausible conjecture as to the value of $W_k(n)$. We begin with

THEOREM 3. *$W_2(n) = [\frac{1}{3}n^2]$, where $[x]$ denotes the integral part of x.*

Proof. Let D be any comma-free dictionary, and define A to be the set of all integers which begin some word of D but never end a word of D. Similarly, let B be the set of integers which both begin and end words of D, and C the set of integers which only end words of D. For example, if $D = \{(43), (41), (35), (25), (15)\}$, then

$$A = \{4, 2\}, \quad B = \{3, 1\}, \quad C = \{5\}.$$

D must evidently consist of words of the forms $(a\ b)$, $(a\ c)$, $(b_1\ b_2)$, or $(b\ c)$, where $a \in A$, $b, b_1, b_2 \in B$, and $c \in C$. But $(b_1 b_2)$ cannot occur, for there is some word in D ending in b_1, and some word beginning with b_2, and the comma-free property therefore excludes $(b_1\ b_2)$. This leaves only words of the forms $(a\ b)$, $(a\ c)$, $(b\ c)$, and it is immediately seen that the set of all these is comma-free. If α is the number of elements of A, β of B, and γ of C, then the number of words in D is at most $\alpha\beta + \beta\gamma + \gamma\alpha$. Maximizing the quantity $\alpha\beta + \beta\gamma + \gamma\alpha$ subject to the constraint $\alpha + \beta + \gamma = n$, we see that α, β, γ should be chosen as nearly equal as possible, in which case

$$\alpha\beta + \beta\gamma + \gamma\alpha = [\tfrac{1}{3}n^2].$$

For example, if $n = 3$ we would take $A = \{1\}$, $B = \{2\}$, $C = \{3\}$, and obtain $D = \{(12), (13), (23)\}$. It is not difficult to see that for arbitrary n we may choose D to be the set of all words w congruent to one of these words (mod 3), where

$$(a_1\ a_2 \ldots a_k) \equiv (b_1\ b_2 \ldots b_k) \pmod{m}$$

means

$$a_j \equiv b_j \pmod{m}, \qquad j = 1, 2, \ldots, k.$$

Thus for $n = 5$, $D = \{(12), (15), (42), (45), (13), (43), (23), (53)\}$.

THEOREM 4. *If k is any even integer, then the upper bound given by Theorem 1 is not attained by $W_k(n)$ provided that $n > 3^{\frac{1}{2}k}$.*

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

Proof. Let $k = 2j$, and let L be a comma-free dictionary. We define S_1 to be the set of all j-tuples $(a_1\, a_2 \ldots a_j)$ which form the first half of some word in L, and S_2 to be the set of k-tuples $(a_{j+1}\, a_{j+2} \ldots a_k)$ which form the second half of some word in L. Then we put

$$A = S_1 \cap S_2', \quad B = S_1 \cap S_2, \quad C = S_1' \cap S_2, \quad D = S_1' \cap S_2',$$

where the prime denotes complementation. The four sets A, B, C, D are mutually exclusive and mutually exhaustive, so that any j-tuple is in one and only one of them. Hence to every k-letter word we may associate a pair (AA), (AB), ... or (DD) depending on which set its first half falls into and which set its second half falls into. As in the proof of Theorem 3, it is seen that for words in L the type (BB) cannot arise, and hence only (AB), (AC), and (BC) remain.

The upper bound of Theorem 1 was the number of primitive equivalence classes. To prove Theorem 4, we will show the existence of a primitive k-letter word, such that no cyclic permutation of it has any of the forms (AB), (AC), or (BC).

Consider the following particular blocks of length j:

$$(1, 1, 1, \ldots, 1, m) \qquad 1 \leqslant m \leqslant n.$$

Let T_i be the cyclic permutation which shifts each letter i units to the left. Define

$$F_m(i) = \begin{cases} 1 \text{ if } T_i\,(1, 1, \ldots, m) \in A \cup D \\ 2 \text{ if } T_i\,(1, 1, \ldots, m) \in B \\ 3 \text{ if } T_i\,(1, 1, \ldots, m) \in C \end{cases}$$

For each m, $F_m(i)$ is a function with a domain of j elements and a range of 3 elements. There can be at most 3^j such functions, and since $n > 3^{\frac{1}{2}k} = 3^j$, there exist two distinct integers p and m such that $F_p \equiv F_m$ for all i. We now claim that no cyclic permutation of the word

$$w = (1\ 1 \ldots p\ 1\ 1 \ldots m)$$

is of the form (AB), (AC), or (BC). For any permutation of w consists of a cyclic permutation of (1 1 ... p) followed by the *same* cyclic permutation of (1 1 ... m) or vice versa. Since $F_p \equiv F_m$, we therefore get only the forms (AA), (AD), (DA), (DD), (BB), or (CC).

In particular, when $k = 4$, Theorem 4 proves that $4W_4(n) < n^4 - n^2$ for $n > 9$. By more delicate arguments it can be shown that this inequality is true for $n \geqslant 5$. On the other hand, if $n = 1, 2, 3$, then $4\ W_4(n) = n^4 - n^2$, as is seen by considering the dictionary D of words $(a\ b\ c\ d)$ satisfying $a < c$, $b \geqslant d$. The question of whether or not $W_4(4) = 60$ is still open. The best that can currently be proved is $W_4(4) \geqslant 56$.

4. Asymptotic Results. In this section we shall prove some theorems about the asymptotic behavior of $W_k(n)$ when k is fixed and $n \to \infty$.

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

THEOREM 5. *The limit*

$$\lim_{n\to\infty} \frac{W_k(n)}{n^k} = \alpha_k$$

exists.

Proof. We shall show that if n_0 is any fixed integer, then

$$\liminf_{n\to\infty} \frac{W_k(n)}{n^k} \geqslant \frac{W_k(n_0)}{n_0^k}.$$

This fact, coupled with the obvious boundedness of the ratio in question, proves the existence of the limit.

Consider then the integer n_0, and let D be a comma-free dictionary containing $W_k(n_0)$ words. For any arbitrary n, form the set S of all words w such that

$$w \equiv w_0 \pmod{n_0},$$

where $w_0 \in D$. (The definition of congruence is given after Theorem 3 together with an example of the present procedure.) S is clearly comma-free, and so if it contains $S_k(n)$ elements, then

$$W_k(n) \geqslant S_k(n).$$

But it is easy to see that

$$\lim_{n\to\infty} \frac{S_k(n)}{n^k} = \frac{W_k(n_0)}{n_0^k}.$$

This completes the proof of the theorem.

THEOREM 6. *If k is odd, then $\alpha_k = 1/k$.*

Proof. By Theorem 1,

$$W_k(n) \leqslant \frac{1}{k} \sum_{d/k} \mu(d)\, n^{k/d}.$$

If k is fixed and $n \to \infty$, the right hand side is asymptotically n^k/k. Hence,

$$\lim_{n\to\infty} \frac{W_k(n)}{n^k} \leqslant \frac{1}{k}.$$

On the other hand, consider the dictionary D defined as follows: Put $k = 2j - 1$ and let D consist of all words $(a_1\, a_2\, \ldots\, a_k)$ such that a_j is greater than any of the other a_i's. D is comma-free, as is easily verified, and the number of elements in D is equal to

$$\sum_{m=1}^{n-1} m^{k-1} \sim \frac{n^k}{k}$$

This shows that

$$\lim_{n\to\infty} \frac{W_k(n)}{n^k} \geqslant \frac{1}{k},$$

and thus establishes Theorem 6.

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

THEOREM 7. *If k is even, then $1/ek < \alpha_k \leqslant 1/k$.*

Proof. The first part of Theorem 6 holds for any fixed k. Hence, $\alpha_k \leqslant 1/k$. To obtain a lower bound, we divide the integers from 1 to n into two disjoint classes U and V. Then let D be the set of all words $(a_1\, a_2 \ldots a_k)$ such that $a_1 \in U$ and $a_2, \ldots, a_k \in V$. D is clearly comma-free, and if the number of elements in V is v, then D contains $(n - v)v^{k-1}$ words. If v could take on all real values, then the maximum of this expression would occur for

$$v = \frac{k-1}{k}\, n\,,$$

and would have the value

$$\frac{n^k}{k}\left(1 - \frac{1}{k}\right)^{k-1}.$$

The fact that v must be an integer has no effect, since taking

$$v = \left[\frac{k-1}{k}\, n\right]$$

gives a lower bound for $W_k(n)$ which is still asymptotically

$$\frac{n^k}{k}\left(1 - \frac{1}{k}\right)^{k-1}.$$

Hence

$$\alpha_k \geqslant \frac{1}{k}\left(1 - \frac{1}{k}\right)^{k-1} > \frac{1}{ek}.$$

For $k = 4$, Theorem 7 gives the bounds

$$\frac{27}{256} \leqslant \alpha_4 \leqslant \frac{1}{4}.$$

A better bound can be obtained from Theorem 5. As shown after Theorem 4, $W_4(3) = 18$, and hence

$$\alpha_4 \geqslant \frac{W_4(3)}{3^4} = \frac{18}{81} = \frac{2}{9}.$$

The exact value of α_k for even k is still an open question.

5. Applications. From their researches in the transfer of genetic information from parent to offspring, Crick, Griffith, and Orgel **(1)** advance the following hypothesis. Genetic information, they suggest, is encoded into a giant molecule (chromosome) by means of an affixed sequence of nucleotides, of which there are four types. Each such sequence is uniquely decodeable into a new protein molecule, consisting of a long sequence of amino acids, of which there are twenty types. They propose that each amino acid is specified by three consecutive nucleotides. However, only twenty of the sixty-four sequences of three nucleotides "make sense." Crick, Griffith, and Orgel

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press

theorize that the twenty sequences of nucleotides actually corresponding to amino acids form a comma-free dictionary. As we have seen, $W_3(4) = 20$, which agrees with the number of amino acids. The reasonableness of this condition can be seen if we think of the sequence of nucleotides as an infinite message, written without punctuation, from which any finite portion must be decodeable into a sequence of amino acids by suitable insertion of commas. If the manner of inserting commas were not unique, genetic chaos could result.

In their search for optimum coding techniques, Shannon, McMillan, and others have studied codes which are uniquely decipherable *in the large*—that is, when the entire message is available. This is a larger class than the comma-free messages, which must be uniquely decipherable *in the small*. In communications applications where only disjointed portions of a message are likely to be received, comma-free codes may indeed be useful. An excellent discussion of codes uniquely decipherable in the large is presented in **(2)**.

References

1. H. C. Crick, J. S. Griffith, and L. E. Orgel, *Codes Without Commas*, Proc. Nat. Acad. Sci., *43* (1957), 416–421.

2. B. McMillan, *Two Inequalities Implied by Unique Decipherability*, IRE Transactions on Information Theory, *2* (1956), 115–116.

3. T. Nagell, *Introduction to Number Theory* (Uppsala, 1951).

https://doi.org/10.4153/CJM-1958-023-9 Published online by Cambridge University Press